

# Selfish Operons: Horizontal Transfer May Drive the Evolution of Gene Clusters

Jeffrey G. Lawrence and John R. Roth

Department of Biology, University of Utah, Salt Lake City, Utah 84112

Manuscript received March 12, 1996

Accepted for publication May 17, 1996

## ABSTRACT

A model is presented whereby the formation of gene clusters in bacteria is mediated by transfer of DNA within and among taxa. Bacterial operons are typically composed of genes whose products contribute to a single function. If this function is subject to weak selection or to long periods with no selection, the contributing genes may accumulate mutations and be lost by genetic drift. From a cell's perspective, once several genes are lost, the function can be restored only if all missing genes were acquired simultaneously by lateral transfer. The probability of transfer of multiple genes increases when genes are physically proximate. From a gene's perspective, horizontal transfer provides a way to escape evolutionary loss by allowing colonization of organisms lacking the encoded functions. Since organisms bearing clustered genes are more likely to act as successful donors, clustered genes would spread among bacterial genomes. The physical proximity of genes may be considered a selfish property of the operon since it affects the probability of successful horizontal transfer but may provide no physiological benefit to the host. This process predicts a mosaic structure of modern genomes in which ancestral chromosomal material is interspersed with novel, horizontally transferred operons providing peripheral metabolic functions.

THE "one gene-one enzyme" hypothesis (BEADLE and TATUM 1941; HOROWITZ and LEUPOLD 1951) and the formation of genetic maps in the 1950s spurred efforts to understand the functional and evolutionary significance of how genes are arranged within the chromosomes of both eukaryotic and prokaryotic taxa. In many organisms, it was discovered that the genes responsible for related, but not identical, functions were frequently located close together on genetic maps.

**Gene clusters are prominent features of bacterial chromosomes:** The most striking examples of gene clusters were found in bacterial taxa (DEMEREK and HARTMAN 1959), such as *Escherichia coli* and *Salmonella typhimurium* (Table 1). In these taxa, the enzymes required for particular biochemical pathways were often found to be encoded by physically proximate genes. Genes, such as the *trp* loci, known to be unlinked in the eukaryote *Neurospora* (BARRATT *et al.* 1954), were found to be clustered in *E. coli* (YANOFSKY and LENNOX 1959). Even more impressive was the tendency of genes for particular pathways to be arranged in the order of their biochemical reactions. The order of the 10 histidine biosynthetic genes within the *E. coli* and *S. typhimurium* his operons is nearly identical to the order of their deduced chemical reactions; the same pattern holds true for the four *E. coli trp* genes. These features were taken as evidence for how gene clusters originated (see discussion of the Natal Model below). The first genetic map of *S. typhimurium* placed 40% of mapped loci into gene clusters (SANDERSON and DEMEREK 1965).

Corresponding author: Jeffrey G. Lawrence, Department of Biology, University of Utah, Salt Lake City, UT 84112.  
E-mail: lawrence@biology.utah.edu

## Gene clusters are rare in eukaryotic chromosomes:

Among eukaryotic taxa, however, genes for related functions were rarely found in close proximity. Putative eukaryotic clusters included the *Aspergillus* adenine, biotin, and proline loci (ROPER 1950; KAUFER 1958), the *Drosophila* bithorax, miniature, yellow, and scute genes (GRÜNBERG 1935; BRIDGES and BREHME 1944; LEWIS 1947; KOMAI 1950; SLATIS and WILLERMET 1953), the *Neurospora* arginine, histidine, pyridoxine, and isoleucine/valine genes (BARRATT *et al.* 1954; MITCHELL 1955; FINCHAM and PATEMAN 1957; NEWMAYER 1957), and mouse developmental genes (DUNN and CASPARI 1945). However, such clusters of genes for related functions are exceptional in eukaryotic genetics. Most genes with associated functions are unlinked, and many of the putative examples of gene clusters have been shown to reflect multiple alleles of a single cistron.

**Models for the origins of gene clusters:** We discuss below three previously suggested general models for the origins of gene clusters: (1) the Natal model, in which gene clusters originate *in situ* by gene duplication and divergence. In this model, gene position is an historical property and provides no direct benefit to the individual. (2) The Fisher model, whereby gene clusters are formed due to selection on coadapted gene complexes, providing a benefit to the individual in the context of a genetically variable, freely recombining population. (3) The Coregulation model, whereby gene clusters facilitate coordinate expression and regulation, providing a selective benefit to the individual. We then describe a new model, the Selfish Operon model, in which gene clusters allow dissemination of functionally related genes via horizontal transfer. In this model,

**TABLE 1**  
**Prominent clusters of biosynthetic and degradative functions identified early during the development of *E. coli* and *S. typhimurium* genetics**

Organism	Locus	Reference
<i>E. coli</i>	<i>ara</i>	LEE and ENGLESBERG (1962)
<i>E. coli</i>	<i>gal</i>	LEDERBERG (1960)
<i>E. coli</i>	<i>lac</i>	LEDERBERG (1962); PARDEE <i>et al.</i> (1959)
<i>E. coli</i>	<i>trp<sup>a</sup></i>	YANOFSKY and LENNOX (1959)
<i>S. typhimurium</i>	<i>his</i>	HARTMAN (1956)
<i>S. typhimurium</i>	<i>ilv</i>	GLANVILLE and DEMEREC (1960)
<i>S. typhimurium</i>	<i>leu</i>	MARGOLIN <i>et al.</i> (1959); GLANVILLE and DEMEREC (1960)
<i>S. typhimurium</i>	<i>pan</i>	DEMEREK <i>et al.</i> (1959)
<i>S. typhimurium</i>	<i>pro</i>	MIYAKE and DEMEREC (1960)
<i>S. typhimurium</i>	<i>thr</i>	GLANVILLE and DEMEREC (1960)
<i>S. typhimurium</i>	<i>try<sup>a</sup></i>	DEMEREK and HARTMAN (1956)

<sup>a</sup>The *try* tryptophan genes were renamed as *trp* genes.

physical proximity provides no selective benefit to the individual organisms, but does enhance the fitness of the gene cluster itself. We suggest that the Selfish Operon model is more likely to explain the evolution of gene clusters in bacteria than other models.

#### THE NATAL MODEL OF GENE CLUSTERING

**Some genes may originate in clusters:** The Natal model postulates that genes are clustered because they were born that way. HOROWITZ (1945, 1965) proposed that synthetic pathways may have evolved in a stepwise fashion, starting with a gene for the last enzyme in a biochemical pathway when a nutrient became limiting in the primitive environment. Each new enzyme would allow another natural compound to serve as a precursor to the nutrient. Additional genes would evolve to augment the pathway as each successive intermediate substrate in the pathway became limiting. LEWIS (1951), extending the ideas of GRÜNBERG (1935), proposed that gene duplication and differentiation could lead to linked loci with related functions (see also STEPHENS 1951). The tendency for the gene order within the *S. typhimurium* *trp* and *his* operons to reflect the order of their corresponding biochemical reactions supported this hypothesis. Since the encoded enzymes would be working on similar substrates, this "assembly line of genes" was viewed favorably (PONTECORVO 1950). In a similar vein, DUNN (1954) proposed that subdivision of a large, multifunctional genetic element could lead to smaller, linked elements with related functions. These theories postulated that the existence of gene clusters reflected the process of gene origin.

As the amino acid sequences of proteins became known, however, these ideas lost merit. Virtually all bacterial operons are composed of genes that show no obvious homology; many are closely related to unlinked genes encoding proteins that catalyze mechanistically similar reactions (*e.g.*, dehydrogenases, kinases, etc.). Few cases of gene duplication and divergence within

an operon have been demonstrated in bacteria (see however FANI *et al.* 1994). On the contrary, the *E. coli* MetB and MetC proteins, which clearly originated by an ancient gene duplication and catalyze successive steps in methionine biosynthesis (BELFAZIA *et al.* 1986), are encoded by unlinked genes.

In contrast to bacteria, the few examples of eukaryotic clusters of functionally related genes appear to be cases of duplication and divergence. For example, the mammalian  $\beta$ -globin gene cluster contains the  $\psi\beta 2$ ,  $\epsilon$ ,  $G\gamma$ ,  $A\gamma$ ,  $\psi\beta 1$ ,  $\delta$ , and  $\beta$  globin genes, all of which arose by duplication and divergence (MANIATIS *et al.* 1980). Four clustered human growth hormone (hGH)/chorionic somatomammotropin (hCS) genes also arose by duplication and divergence (JONES *et al.* 1995). Therefore, while the Natal model is likely to account for the few gene clusters among eukaryotes, it appears unlikely to explain the more extensive gene clusters found in bacteria.

**The Natal model cannot explain the persistence of gene clusters:** Not all early ideas on the significance of gene clusters focused on the origin of the component genes. DEMEREK and HARTMAN (1956) postulated that, regardless of how gene clusters originated, natural selection must act to prevent their separation. It followed, then, that natural selection might have worked to aggregate previously separated loci. DEMEREK and HARTMAN (1959) noted that the "mere existence of such arrangements shows that they must be beneficial, conferring an evolutionary advantage on individuals and populations which exhibit them." Few explanations were offered as to what such a benefit could be. HARTMAN (1956) proposed a "position effect," in that the arrangement of the genes provided the cell with a selective advantage not conferred by unlinked genes; no mechanistic basis for this advantage was suggested. DEMEREK and DEMEREK (1956) were more specific in outlining a "position effect" and proposed that the biochemical reactions were localized to the genes in prokaryotes and

were relegated to extranuclear sites in eukaryotes. Therefore, eukaryotic taxa would have lost the selection for gene clustering, and previously clustered genes had dispersed.

In an alternative model, HARTMAN *et al.* (1960) proposed that gene clustering provided no biochemical benefit to the cell. Rather, the close proximity of the genes allowed a single transducing particle to repair multiple lesions in the genes for a single biochemical pathway. This model provided a selection for restraining the separation of genes originating as clusters. However, it correspondingly provided selection against the clustering of dispersed genes; any chromosomal rearrangement bringing genes together would preclude repair by transduction from a nonrearranged strain.

**Diverse operons contain homologous genes:** As the primary sequences of proteins accumulated, the comparison of homologous proteins from distantly related organisms revealed the relationships among the genes (ZUCKERKANDL and PAULING 1962). Families of globins (NOLAN and MARGOLASH 1968), cytochromes (SMITH and MARGOLASH 1964), and other proteins were deduced. These families did not illuminate the question of bacterial operon origins until families of functionally distinct dehydrogenases were inferred. ROSSMAN and colleagues (BUEHNER *et al.* 1973; ROSSMAN *et al.* 1974) proposed that the NAD-binding proteins lactate dehydrogenase, malate dehydrogenase, alcohol dehydrogenase, and glyceraldehyde-3-phosphate dehydrogenase were derived from a common ancestor. In addition, FAD-binding proteins, such as flavodoxin, were proposed to be distantly related to these NAD-binding proteins. Since dehydrogenases are found in many different operons, these data strongly suggested that operons were composed of genes that arose independently, from distinct ancestors (unrelated to each other), and were later assembled into clusters. Membership in a gene family has since become an established method for discerning the function of a novel protein (ORENGO *et al.* 1993; PETRILLI 1993; HOLM and SANDER 1994).

#### THE FISHER MODEL OF GENE CLUSTERING

**Genes clusters may reflect coadaptation:** The Fisher model postulates that genes cluster if specific alleles work well together. Before the gene clustering debate began, FISHER (1930) noted that when specific alleles of two genes worked well together, the deduced linkage of the two genes would increase. This increase in the observed linkage resulted from selection for specific genotypes (*e.g.*, "AB" and "ab", where "A" and "a" are alleles at one locus and "B" and "b" are alleles at another) and counterselection against recombinants (*e.g.*, "Ab" and "aB"). This idea was extended by several workers (BODMER and PARSONS 1962; STAHL and MURRAY 1966) who suggested that such selection could

lead to the physical clustering of genes. The increased physical proximity would reduce the frequency of recombination events that disrupt coadapted loci. In eukaryotes, the predictions of this model are not seen since there are few examples of clustered, functionally related genes.

The Fisher model has been cited as motivating the clustering of genes within bacteriophage genomes (STAHL and MURRAY 1966; BOTSTEIN 1980; CAMPBELL and BOTSTEIN 1983; CASJENS *et al.* 1992). In both lambdaoid and T4 family bacteriophages, genes encoding proteins that function together as a logical group (*e.g.*, head or tail proteins) are found in clusters. According to the Fisher model, this arrangement would allow recombination between lambdaoid bacteriophages to generate new combinations of logical groups but would not disrupt the individual clusters of genes whose products must work together most intimately. This theory has been termed the "module" approach to bacteriophage evolution (BOTSTEIN 1980; CAMPBELL and BOTSTEIN 1983). As predicted by the Fisher model, the genes within many bacteriophage clusters encode proteins that interact physically; clustering minimizes the distance between them (CASJENS 1974; CASJENS and HENDRIX 1988; CASJENS *et al.* 1992).

**The Fisher model requires frequent recombination:** The Fisher model requires two conditions to provide selection for gene clustering. First, there must be sufficient genetic variation at the loci under selection so that multiple coadapted gene complexes (AB and ab) may arise. Second, there must be sufficient recombination so that the coadapted allele combinations are regularly disrupted. It is this potential for disruption that selects for clusters. In eukaryotes, such recombination is provided by meiosis and sexual reproduction, but the source of abundant recombination is less evident for bacteria and bacteriophage, which reproduce asexually. It is particularly hard to envision how the Fisher model might drive the clustering of genes for bacterial metabolic processes. Separate enzymes in a metabolic pathway, if they physically interact at all, do so to a lesser degree than that seen for structural proteins. MARTIN *et al.* (1971) failed to detect any co-association of enzymatic function among isolated histidine biosynthetic enzymes. In a more sensitive screen, when suppressors of missense mutations in the *hisD* gene (encoding the enzyme acting last in the histidine biosynthetic pathway) were isolated, none were allele-specific suppressors mapping to other genes in the *his* operon (J. Bullock and J. R. Roth, unpublished results). It is less likely that alleles of metabolic genes are coadapted to work with particular alleles of other genes in the same pathway. In addition, the asexual nature of bacterial reproduction does not obligate the large-scale recombination required to disrupt the nascent coadapted gene complexes. Most bacterial species exhibit population structures indicative of little recombination among conspe-

TABLE 2  
Functionally related unclustered genes identified early during the development of *E. coli* and *S. typhimurium* genetics

Organism	Locus	Linkage groups	Reference
<i>E. coli</i>	<i>arg</i>	5	GORINI <i>et al.</i> (1962)
<i>E. coli</i>	<i>pyr</i>	3	BECKWITH <i>et al.</i> (1962)
<i>S. typhimurium</i>	<i>ade<sup>a</sup></i>	5	YURA (1956)
<i>S. typhimurium</i>	<i>cys</i>	4	DEMEREK <i>et al.</i> (1955); HOWARTH (1958)
<i>S. typhimurium</i>	<i>met</i>	4	SMITH (1961)

<sup>a</sup>The *ade* genes (adenine-requiring) were renamed *pur* (purine) genes.

cific strains (DESJARDINS *et al.* 1995; MAYNARD SMITH *et al.* 1993).

#### THE COREGULATION MODEL OF GENE CLUSTERING

**Gene clusters can be regulated efficiently:** The operon model for coregulation of genes under a common control mechanism stimulated new ideas to explain gene clustering (PARDEE *et al.* 1958; JACOB *et al.* 1960; JACOB and MONOD 1962). Genes found together, they explained, could be induced and repressed simultaneously by control at a single site, termed an operator. This control offered a selective mechanism by which a cluster of genes could provide a selective advantage over the same genes at dispersed sites. The operon offered both economy of expression and fixed relative product abundance to genes expressed from a single promoter. This model provided a rationale for both coregulation of genes and their clustering. As a result, the concept of the operon as the causative agent in gene clustering was widely accepted as a general explanation of why bacterial chromosomes were organized into gene clusters (AMES and MARTIN 1964). The force of this model was mitigated by the discovery of coregulated, unlinked genes (Table 2). These cases showed that clustering is not a prerequisite for coregulation.

**Coregulation cannot drive gene clustering:** A more serious problem of the Coregulation model was its failure to suggest a plausible series of intermediate steps in the evolution of gene clusters. The regulatory benefits of an operon are derived from the cotranscription of multiple genes from a single promoter. Without cotranscription, genes 500 bases or 500 kb apart are, in effect, equally distant if transcription termination sites are located between them. No benefit is derived from proximity until cotranscription is possible.

If the only selective value of gene clustering were the final operon, the process of operon coalescence would have to occur in a single step, placing previously unlinked genes under the control of a single, regulated promoter. In effect, the extraordinarily rare event of a chromosomal rearrangement precisely juxtaposing two related genes would have to be strongly selected, so that it when it occurs it has a high likelihood of rising to high frequency

in the population. Moreover, such an event must occur for each gene added to every operon observed. Rearrangements altering gene orders include inversions, which are relatively rare (ROTH *et al.* 1996), duplications, which are common and yield novel join points but are unstable, deletions, which are stable but permanently eliminate intervening DNA that may be selectively valuable, and transpositions, in which mobile genetic elements support the rearrangement of gene order.

A paradox is evident if coregulation were to drive gene clustering. The regulatory benefit of placing two genes under the control of a single promoter must be strong to ensure that the very rare rearrangement precisely juxtaposing two related genes is fixed in the population before being lost by stochastic processes. However, such strong selection is unlikely to be conferred by an allowing an operator to regulate one additional gene of a pathway. Therefore, the well-regulated promoter responsible for driving gene clustering cannot provide maximum benefit until all of the genes are clustered. Alternatively, well-regulated operators may slowly be selected simultaneously at unlinked loci, as is seen with the *E. coli arg*, *met*, and *pur* genes. Once this occurs, there is no selection for gene clustering.

**Potential advantages of cotranscription are not exploited:** An additional benefit of cotranscription is the ability to produce proteins in equimolar amounts. However, due to the different catabolic efficiencies of different enzymes, it is unlikely that the precisely equimolar amounts which might be produced from a single transcript would, in fact, be beneficial. As expected, genes within a single operon often show different translational efficiencies (VAN DE GUCHTE *et al.* 1991) and different mRNA half-lives (KEPES 1967; BLUNDELL *et al.* 1972), which contribute to the nonuniform levels of proteins encoded by a single operon. For example, the first enzyme encoded by the *S. typhimurium his* operon, HisG, is present at four times the concentration of the adjacent HisD or HisC proteins (WHITFIELD *et al.* 1970). Even in cases where final molar ratios can be precisely predicted, such as for ribosomal proteins, cotranscribed genes produce different levels of proteins. The ribosomal proteins L1, L10, L11, and L7/L12 are cotranscribed from an operon at minute 90 of the *E. coli*

chromosome (BRÜCKNER and MATZURA 1981). However, the L7/L12 protein is present in four copies per 50S subunit, while the other three proteins are present in one copy each (HARDY 1975; SUBRAMANIAN 1975). We propose that coregulation may be a benefit derived secondarily from operon formation, and may provide a selective influence for maintaining operon organization. However, we feel that coregulation is unlikely to provide sufficient selection to drive the clustering of dispersed genes. Hence, the maintenance of operons may include selective forces not contributing to the origin of the clusters.

#### THE SELFISH OPERON MODEL OF GENE CLUSTERING

**Genes move by horizontal transfer:** To explain the formation of gene clusters in bacteria, we offer a model that relies upon the horizontal transfer of DNA between organisms. Inheritance of genetic information in bacteria occurs primarily by vertical transfer; that is, transfer from a parent cell to daughter cells at cell division. However, the transfer of DNA between organisms independent of reproduction is known to occur; this process has been termed horizontal transfer [reviewed by KIDWELL (1993) and SYVANEN (1994)] and entails the transfer of DNA between species. Although examples of horizontal transfer involving eukaryotic taxa have been documented (MOURANT 1971; BANNISTER and PARKER 1985; BRISSON-NOEL *et al.* 1988; HILDEBRANDT *et al.* 1989; DOOLITTLE *et al.* 1990) they are, for the most part, considered isolated events. In contrast, horizontal transfer among prokaryotes is mediated by common processes [(see review by SYVANEN (1994)], including transducing bacteriophages, conjugative plasmids, or the direct uptake of foreign DNA. We define horizontal transfer and the mobilization of DNA between bacterial species. The transfer of genetic information among conspecific strains is typically denoted as recombination among bacterial isolates.

**Genes for weakly selected functions can be lost:** Bacterial genomes encode both critical functions, required for central metabolic processes, and merely useful functions, which provide sporadic benefits but are not continuously essential. Such noncritical functions include the degradation of unusual compounds as sources of carbon and energy; these compounds may not always be present or may represent a minor fraction of available growth substrates. We term such functions "weakly selected". Weakly selected functions include those employed frequently for relatively unimportant tasks or those employed only under specific, rarely encountered environmental conditions. Natural selection on genes encoding rarely used proteins may be temporally or spatially heterogeneous. During periods of relaxed selection, such genes may accumulate base substitutions rapidly since natural selection would not act to

remove null alleles from the population. Consequently, genomes with multiple defects in a single function can rise to high frequency by genetic drift.

A simple calculation demonstrates this process of loss of gene function. If the coefficient of selection ( $s$ ) of a mutation is sufficiently small, the mutation may be considered effectively neutral (KIMURA 1983). In haploid populations, the magnitude of an effectively neutral selection coefficient may be approximated as  $s \leq 1/2N_e$ , where  $N_e$  is the effective population size. Since natural selection may be temporally or spatially heterogeneous, we define a weakly selected function in terms of an average selective coefficient,  $\bar{s}$ , for that allele over all environments. If  $\bar{s} \leq 1/2N_e$ , then the null alleles at that locus are effectively neutral. Every generation, a total of  $N_e\mu$  mutations will arise (at a frequency  $\mu$  per cell in a population with effective size  $N_e$ ). Since each mutation has a probability of  $1/N_e$  to sweep the population, a total of  $N_e\mu \times 1/N_e = \mu$  mutations will be fixed per generation. Therefore, effectively neutral null mutations may sweep a population in an average of  $1/\mu$  generations. For example, consider a function requiring five genes (or 5000 bp) in *E. coli*. If the probability of mutation is  $10^{-9}$  per base pair, then the mutation rate can be estimated as  $\mu = 5 \times 10^{-5}$ /generation for these genes. Neutral alleles will sweep the population in 200,000 generations or  $\sim 1000$  years. (This frequency does not consider other factors, such as the loss by spontaneous deletion, or the effects of null alleles hitchhiking on linked, selectively advantageous mutations at other loci.)

Consider the hypothetical *wsf* locus encoding a weakly selected function (Figure 1). If natural selection is relaxed, null alleles (*wsf*<sup>-</sup>) will eventually dominate the population and the weakly selected function will be lost. In other words, *wsf*<sup>+</sup> cells have an insufficient selective advantage over *wsf*<sup>-</sup> cells to prevent their eventual loss by genetic drift. If more than one gene is required for this function, the target for potential mutations is correspondingly larger, and the rate of stochastic loss of this function is correspondingly higher. Once one gene contributing to a weakly selected, multigene function is lost, all selection is removed from the remaining genes and mutants with multiple defects rapidly arise.

Moreover, cells with multiple defects in a single metabolic pathway may be selected over cells with single lesions. In many cases, single mutations may confer a selective disadvantage, for example, by allowing the buildup of toxic metabolic intermediates. In such cases, additional mutations in the same pathway would provide a selective advantage. Therefore, cells multiply mutant in particular pathways may show a selective advantage over singly mutant cells. This was observed for some adenine mutants of *Neurospora* (MITCHELL and MITCHELL 1950); double mutants held a selective advantage over single mutants.

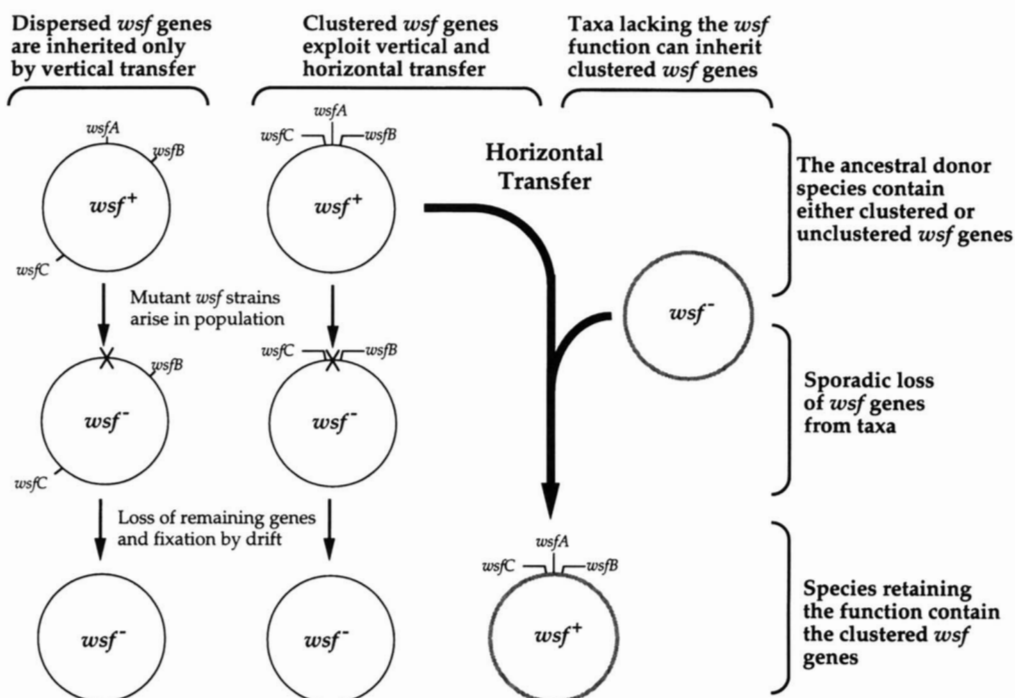


FIGURE 1.—Model for the transfer of gene clusters. Circles represent bacterial chromosomes; *wsfABC* denotes genes for a weakly selected function. Corresponding *wsf* phenotypes are provided at the center of each chromosome.

**Horizontal transfer allows genes for weakly selected functions to escape extinction:** The potential loss of the *wsf* genes from their native species may not mean that the *wsf* genes are doomed to extinction. Horizontal transfer may have mobilized the *wsf* genes to a recipient taxon before their stochastic loss from the donor taxon. Hence, horizontal transfer allows the persistence of genes that could be doomed to evolutionary extinction if vertical transfer were their only means of inheritance. Such mechanisms are thought to influence the evolution of transposons in eukaryotes, in which autonomous transposons must be horizontally transferred regularly to select for transposition function (HARTL *et al.* 1992; HURST *et al.* 1992). Unlike transposons, however, more than one *wsf* gene may be required to perform the weakly selected function. If so, the lateral transfer of only one *wsf* gene to a multiply deficient recipient would not provide the function to the recipient cell. Acquired genes that do not confer a selective advantage would not rise to high frequencies in the population; rather, they would be again lost from bacterial populations by deletion or accumulation of mutations. Only when all *wsf* genes required for the novel function are transferred at once will a potentially beneficial phenotype result and a selective benefit be realized (Figure 1). This process of loss and reacquisition can occur both within and among bacterial species; for the purposes of this paper, we will concentrate on the transfer of Selfish Operons among bacterial species.

**The cluster is a selfish property of the constituent genes:** Consider three genes, *wsfA*, *wsfB*, and *wsfC*, required for a weakly selected function. If these genes are

scattered on a chromosome, they may be propagated only by vertical transfer. Eventually, they will be lost by the accumulation of null mutations as described above (Figure 1). If these genes are clustered, however, they may be propagated by both vertical and horizontal transfer. These genes can escape loss by genetic drift by exploiting transfer to novel genomes. Only when all three genes are transferred is the selective benefit to the recipient cell realized. The physical proximity of *wsf* genes provides no selective benefit to donor organism; organisms with clustered or unclustered *wsf* genes are equally fit. However, physical proximity provides a strong advantage to the *wsf* genes themselves when competing via lateral transmission with unclustered alternative alleles. Therefore, the gene cluster can be considered a selfish property. The cluster is advantageous only to the genes themselves not to the immediate host organism. This feature distinguishes the Selfish Operon model from the Coregulation model of gene clustering in which the host gains a fitness advantage due to better regulation. This feature also distinguishes the Selfish Operon model from the Fisher model in which the physical proximity of coadapted alleles increases the organism's fecundity by reducing the frequency of less fit recombinant offspring.

**Bacteriophage genes—the Fisher model revisited:** Bacteriophage genomes contain tight clusters of genes encoding highly coadapted proteins; these proteins physically interact to form bacteriophage head or tail structures. Since these proteins are coadapted, investigators have extended the Fisher model to explain the clustering of these genes (see above). According to the

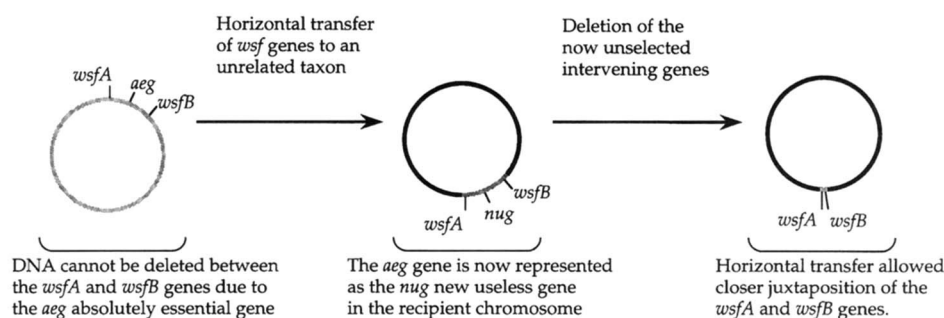


FIGURE 2.—Rapid clustering of genes within foreign genomes. Circles represent bacterial chromosomes; *wsfAB* denotes genes for a weakly selected function, *aeg*, absolutely essential gene, *nug*, now useless gene.

extended Fisher model, recombination among diverse bacteriophages disrupts coadapted gene complexes, and the recombinants are counterselected. As a result, bacteriophages are proposed to evolve tight clusters of coadapted genes; these clusters minimize detrimental recombination between coadapted genes. Yet recombination among bacteriophages is currently viewed as a very rare event between distantly related partners; therefore, extensions of the Fisher model cannot explain gene clustering in bacteriophages. We believe that the evolution of these gene clusters is explained better by the Selfish Operon model.

Since the genes for bacteriophage proteins are coadapted, successful propagation of any one gene requires that all coadapted genes are transferred in a single recombination event. Transfer of a single gene of a coadapted complex to a bacteriophage genome lacking the analogue of this gene would be unsuccessful; since the remaining genes in the recipient genome are coadapted, any newly acquired single gene would not function properly in this new context. In contrast, the introduction of an entire, coadapted gene complex into a bacteriophage genome lacking one gene of its analogous complex would be successful. Moreover, the products of the remaining genes of the recipient's complex could not interact with the products of the newly introduced, coadapted gene complex and therefore would be removed from selection and lost. Hence, the coadapted nature of the genes can accelerate the co-transfer of genes contributing to a single function. The Selfish Operon model predicts that clusters of functionally related genes can colonize naïve genomes or genomes having lost two genes of a single pathway. Clusters of coadapted, functionally related genes can invade genomes from which only a single gene has been lost.

Rather than disrupting coadapted complexes, as in the Fisher model, recombination in the Selfish Operon model facilitates the propagation of clusters of coadapted genes. In contrast to the Fisher model, the transfer of genes between bacteriophage genomes can drive clustering even if recombination is rare. Since the recombination events generally do not involve closely related bacteriophage genomes, they can be thought of as horizontal transfer events mobilizing small sequence elements. Viewed in this manner, the Selfish Operon

model explains the clustering of bacteriophage genes and bacterial chromosomal genes by parallel pathways.

#### PROPERTIES OF THE SELFISH OPERON MODEL

**Horizontal transfer accelerates gene clustering:** The Selfish Operon model allows genes to cluster into an operon by a series of approximations. This is an attractive alternative to the "instant operon" required by the Coregulation model. The efficiency by which multiple genes are cotransferred increases as genes are brought closer together. Therefore, genes can be slowly moved into clusters even before coregulation is possible. Any rearrangement that brings two or more genes with co-operating products closer together increases that group's ability to be mobilized. In contrast, the Coregulation model requires that operons be formed in one step (see above).

Horizontal transfer can contribute more directly to the clustering of genes by making the intervening DNA nonessential. Following horizontal transfer, an introgressed DNA fragment containing loosely clustered genes will be foreign to the host. Since this is foreign DNA, the intervening material (between the selected, loosely clustered genes) will not be essential for the growth of the recipient cell. This intervening DNA is subject to spontaneous deletions, which can bring the loosely clustered genes into closer proximity (DEMEREK 1960) (Figure 2). In this manner, loosely clustered genes transferred horizontally may be brought rapidly and incrementally into very close proximity. In this model, the deletions, which juxtapose genes, delete foreign DNA, not essential DNA. The *wsf* genes contained in the horizontally transferred DNA will be selected; the intervening material is unselected and can be lost by deletion. If the intervening material encodes products that disrupt the metabolism of the recipient cell, deletion of these sequences may be selected, thereby accelerating further the clustering of the beneficial *wsf* genes. In contrast, in the Coregulation and Fisher models such deletions are likely to remove selectively valuable genes from the native chromosome. Therefore, horizontal transfer not only selects for previously clustered genes, it actively participates in the process of bringing genes progressively closer together by reduc-

ing the constraints on deletion of the intervening material.

### **Selfish operons and the evolution of promiscuity:**

The evolution of gene clusters by the Selfish Operon model does not require cotranscription. Yet bacterial chromosomes are notable for clusters of genes under the control of single promoters, that is, for operons. If each gene of a transferred cluster must be transcribed by a separate promoter, it is likely that one or more of these promoter sequences may not function in some recipient genomes which recognize different promoter types. If so, the gene cluster would not provide a selectable phenotype and would be lost by deletion.

*Cotranscription is a selfish property of genes:* We suggest that the cotranscription of genes for specific metabolic functions may facilitate horizontal transfer of gene clusters into genomes with RNA polymerases that recognize different promoter sites than those found in the donor genome. An introgressed operon requires only a single new promoter sequence, which may be provided by the recipient at the site of insertion. Therefore, while physical proximity is strongly selected so that cotransfer may occur, cotranscription may be subsequently selected to allow transcription of all genes in the widest possible variety of new environments. Indeed, the new host may provide the single promoter that allows expression of the laterally transferred genes; the individual promoters for the donor-cell RNA polymerase may be superseded by a single promoter site well recognized by the recipient-cell RNA polymerase. The cotranscription of genes within operons may be a trivial result of the inevitable loss of individual, species-specific promoter sequences.

*Cotranscription may select for the maintenance of gene clusters:* Only following operon formation could regulation at a single operator provide a selective advantage to the cell. However, once an operon is formed, coregulation may provide a selection against the subsequent dispersion of genes. Once genes have clustered into an operon (for selfish purposes), the regulation of the single promoter may provide selective benefits to the host organism. In this manner, coregulation may provide a selection for the maintenance of operons, even though it cannot provide a selection for the formation of operons. Moreover, the close proximity of genes within an operon reduces the probability of gene dispersal, since the chromosomal rearrangements leading to operon disruption are likely to damage some of the genes within the operon. The relative contributions made by the Coregulation model and the Selfish Operon model toward the maintenance of gene clusters cannot be easily estimated and would depend on the selective benefit of coregulation enjoyed by individual operons.

*Translational coupling may be a selfish property of operons:* In a fashion similar to transcription initiation, translation initiation requires sequence signals that vary among species. Therefore, ribosomes of a foreign host

may not recognize all of the translation initiation sites in a transferred operon. Within bacterial operons, the translation of downstream genes is sometimes initiated by ribosomes completing the translation of upstream genes; this has been termed "translational coupling" (OPPENHEIM and YANOFSKY 1980). The *de novo* initiation of translation is not required for the second gene of a coupled pair; rather, ribosomes completing the translation of the first gene do not dissociate fully from the mRNA. After dissociation of the 50S subunit, the 30S subunit may reinitiate translation at a physically proximate translation start site (MARTIN and WEBSTER 1975). We view translational coupling as a mechanism to ensure translation by foreign ribosomes of all proteins encoded by a single message. Hence, five clustered genes not organized into an operon would require five transcription initiation and five translation initiation events for expression; an operon of five translationally coupled genes would require a single transcription initiation and a single *de novo* translation initiation event. While translational coupling would offer this advantage to horizontally transferred genes, it is not obligatory.

*The clustering of operons with genes encoding their trans-acting regulators is selfish:* A notable feature of some bacterial operons is the adjacent location of a gene encoding a *trans*-acting regulatory protein. In *E. coli*, this arrangement is observed for the separately transcribed *putA* and *putP* genes, *araC* and *araBAD* genes, *rhaRS* and *rhaBAD* genes, *ebgR* and *ebgACB* genes, *melR* and *melAB* genes, and *lacI* and *lacZYA* genes. In each case, the genes listed first encode a *trans*-acting regulatory protein; the gene(s) listed second comprise a single transcription unit controlled by that regulatory protein. The close proximity of the regulatory genes and the regulated operons is not essential for the control mechanism since these proteins can act at distant sites. For example, the AraC regulatory protein controls both the *araBAD* operon (linked to the *araC* gene) and the distant *araEFG* operon; thus it functions effectively in *trans*, suggesting that the proximity of the *araC* gene and the *araBAD* operon is not necessary for function. Similarly, the chromosomally encoded LacI protein routinely regulates *lac* promoter sequences on common cloning vectors and could thus act in *trans* on an unlinked, chromosomal *lacZYA* operon. Although the proximity of these regulatory genes to their targets cannot be explained simply on a functional basis, their proximity can be explained by the Selfish Operon model. The adjacent location of the regulatory gene and the regulated operon may have been selected since that proximity has allowed efficient cotransfer of the operon and its regulatory apparatus.

**The *S. typhimurium* *cob* operon, an example of loss and reacquisition:** We believe that the organization of the cobalamin (coenzyme B<sub>12</sub>) biosynthetic operon and the propanediol degradation operon of *S. typhimurium* exemplifies the predictions the Selfish Operon model.

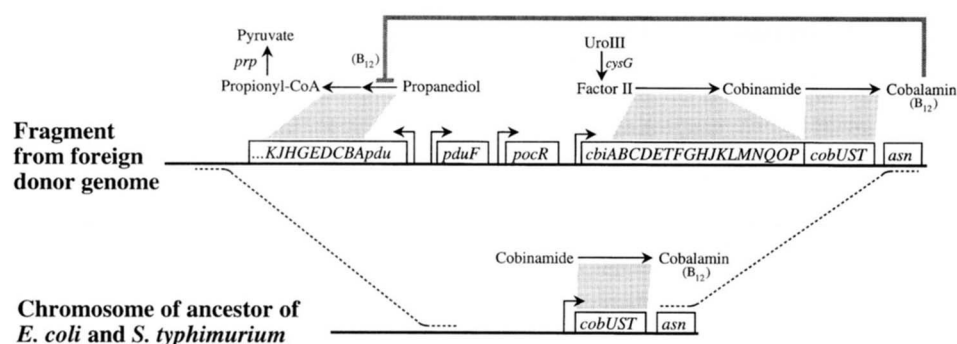


FIGURE 3.—Schematic of the horizontal transfer event introducing the *cob* and *pdu* operons into the *Salmonella* genome. Genes and operons are represented by boxes; grey lines demarcate reactions catalyzed by the encoded proteins. The dotted lines represent the proposed sites of introgression of the *cob* and *pdu* operons into the *Salmonella* chromosome.

*S. typhimurium* synthesizes cobalamin, employing the 20 genes of the *cob* operon, only under anaerobic growth conditions (JETER *et al.* 1984). Propanediol degradation depends upon B<sub>12</sub> as a cofactor and requires the enzymes encoded by the *pdu* operon, transcribed divergently from the *cob* operon (Figure 3). Since the *cob* and *pdu* operons are both induced by propanediol (BOBIK *et al.* 1992), the degradation of propanediol is believed to provide the primary selection for cobalamin biosynthesis in *S. typhimurium*. These two functions, requiring nearly 1% of the *S. typhimurium* chromosome, must be under strong selection in this genus; *Salmonella* species are almost universally capable of propanediol degradation and cobalamin biosynthesis (LAWRENCE and ROTH 1996). The functions of the *cob* and *pdu* operons are the basis of metabolic tests designed to discriminate between *Salmonella* spp. and other enteric bacteria (RAMBACH 1990).

Most enteric bacteria synthesize cobalamin under aerobic conditions and employ the cofactor in glycerol and propanediol dehydratases (LAWRENCE and ROTH 1996). The immediate ancestor of *Salmonella* spp. and *E. coli* is believed to have lost both cobalamin synthesis and the two dehydratases. *E. coli* isolates reflect these losses and neither synthesize cobalamin *de novo*, nor degrade propanediol or glycerol in a cobalamin-dependent fashion (LAWRENCE and ROTH 1995, 1996). In contrast, *Salmonella* spp. appear to have acquired a foreign cluster of genes encoding proteins for cobalamin synthesis (*cob*) and propanediol degradation (*pdu*). Since these operons are adjacent, it is likely that both were acquired from a single transferred fragment; this event is diagrammed in Figure 3 and evidence for the horizontal transfer event is detailed elsewhere (LAWRENCE and ROTH 1995, 1996). It is likely that the simultaneous introduction of the adjacent *cob* and *pdu* operons into the ancestral genome of *Salmonella* allowed that genome to rise to high frequency since it introduced a new degradative pathway and the biosynthetic pathway for the required cofactor.

The *cob* and *pdu* operon region contains all necessary proteins for the synthesis of cobalamin and its use in the degradation of propanediol. The precursors to cobalamin biosynthesis and the product of propanediol degradation are constituents of existing *Salmonella* me-

tabolism. The substrate for cobalamin biosynthesis is a methylated tetrapyrrole that is produced by the CysG protein during siroheme synthesis (SPENSER *et al.* 1993; FAZZIO and ROTH 1996). Siroheme is required for cysteine biosynthesis, and the CysG protein is expressed constitutively at a basal level. Therefore, the substrate for the Cob biosynthetic proteins is consistently available in the *Salmonella* cellular environment. Similarly, the product of propanediol degradation (propionyl-CoA) may be converted to pyruvate by the enzymes involved in propionate degradation (J. TITTENSOR and J. R. ROTH, unpublished data). Hence, the product of propanediol degradation by the Pdu enzymes readily enters *Salmonella* central metabolism. The *trans*-acting *pocR* regulatory gene, as expected, is located between the *cob* and *pdu* operons (see above and Figure 3).

This arrangement of genes in the *cob/pdu* operon cluster includes hierarchical levels of selfish gene clusters. The 20 cotranscribed *cob* genes can provide cobalamin synthesis when mobilized to recipient genomes (Figure 3). Similarly, the adjacent *pdu* operon can be transferred into foreign genomes to confer the ability to degrade propanediol (Figure 3). Together, the adjacent *cob* and *pdu* operons together form a selfish regulon, providing the functions of propanediol degradation and synthesis of the cofactor required for that process.

This process of loss and reacquisition has also been proposed to account for the remarkable similarity between the structure and sequence of the *trp* operon in *Brevibacterium lactofermentum* and the *trp* operons of enteric bacteria (MATSUI *et al.* 1986; CRAWFORD and MILKMAN 1991). CRAWFORD and MILKMAN (1991) postulated that the *trp* functions were lost from the ancestor of *B. lactofermentum*, and only the transfer of the entire *trp* operon could have restored the tryptophan biosynthetic functions.

**Introgressed operons are common among *E. coli* and *S. typhimurium*:** The *cob* and *pdu* operons demonstrate that complex functions may be gained by an organism by horizontal transfer. Many such introgressed operons have been identified within the *E. coli* and *S. typhimurium* chromosomes (Table 3). The *rfb* locus encodes enzymes necessary for the synthesis of the O antigen polysaccharides of enteric bacteria. Reeves and cowork-

TABLE 3  
Operons of exogenous origin in the *E. coli*  
and *S. typhimurium* genomes

Organism	Locus	Map position	Reference
<i>E. coli</i>	<i>lac</i>	8	BUVINGER <i>et al.</i> (1984)
<i>E. coli</i>	<i>rfa</i>	81	KLENA <i>et al.</i> (1993)
<i>E. coli</i>	<i>rfb</i>	44	STEVENSON <i>et al.</i> (1994)
<i>S. typhimurium</i>	<i>cob</i>	42	LAWRENCE and ROTH (1995; 1996)
<i>S. typhimurium</i>	<i>rfb</i>	45	REEVES (1993)
<i>S. typhimurium</i>	<i>oad</i>	?	WOEHLKE <i>et al.</i> (1992)
<i>S. typhimurium</i>	<i>spa</i>	57	GROISMAN and OCHMAN (1993)

ers (REEVES 1993; XIANG *et al.* 1994) have determined that the *rfb* locus of *Salmonella* spp. has been introduced by horizontal transfer. Moreover, this 15 gene operon appears to be an evolutionary mosaic, representing numerous horizontal transfer events that assembled the *rfb* locus from many different genomes. The *rfb* and *rfa* loci of *E. coli* also reflect patchwork patterns of gene composition indicating different evolutionary origins for gene subclusters (KLENA *et al.* 1993; LIU and REEVES 1994; STEVENSON *et al.* 1994).

The *spa* genes contribute to the ability of *Salmonella* spp. to invade eukaryotic cells. This operon is also purported to be of exogenous origin; *spa* homologues are not found among closely related taxa and the GC contents of the genes (~30–40% G+C) are atypical of *S. typhimurium* coding sequences (GROISMAN and OCHMAN 1993). The mosaic pattern of GC content among *spa* genes suggests that, like the *rfa* and *rfb* operons, the *spa* operon may represent another example of an operon of genes assembled from different chromosomes. Like the *spa* operon, genes of the *S. typhimurium oad* operon have aberrant GC contents (~65% G+C and ≤87% G+C in the third codon position) and other DNA sequence features indicative of horizontal transfer (OCHMAN and LAWRENCE 1996). Similarly, the *cat* operon of *Acinetobacter calcoaceticus* has a mosaic structure indicative of recent assembly from multiple sources (SHANLEY *et al.* 1994). The nucleotide composition of genes are commonly used as indicators of possible exogenous origin (MÉDIGUE *et al.* 1991; WHITTAM and AKE 1992; OCHMAN and LAWRENCE 1996); the *rfa* (35–39% GC), *rfb* (31–40% GC), *cob* (59% GC), and *pdu* (59% GC) operons described above all show GC contents atypical of the *S. typhimurium* genome.

**Simulations of the Selfish Operon model:** To test mathematically the predictions of the Selfish Operon model, a computer model was developed (Figure 4). A collection of virtual taxa was created bearing ≤10 genes required for a hypothetical function. In any particular simulation, all taxa in the collection had the same number of genes contributing to this function; the genes

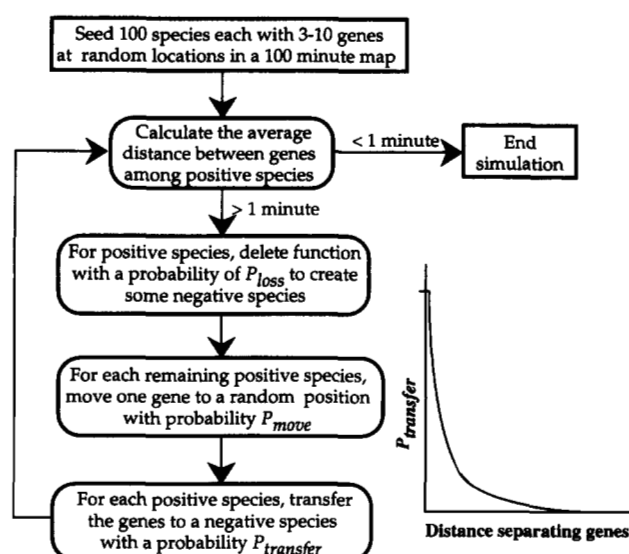


FIGURE 4.—Flow chart describing computer simulation of horizontal transfer events. The numbers of positive taxa were allowed to vary between 10 and 900 species. The distance between loci was calculated as the minimum chromosome arc containing all genes in the putative cluster. Average distance was calculated as the arithmetic mean of these distances among positive taxa. Positive taxa lose the simulated function with a probability of  $P_{loss}$  per cycle. For each positive taxon, one gene may translocate to a random location within the 100-min chromosome with a probability of  $P_{move}$  per cycle. For each positive taxon, the genes may be transferred to a negative taxon with a probability of  $P_{transfer}$  per cycle;  $P_{transfer}$  varies inversely with the distance between the loci in that taxon; a maximal  $P_{transfer}$  occurs when 0 min separate the loci.

were placed randomly on a linear, 100-min genetic map. Taxa carrying functional alleles of all genes are termed “positive”; taxa lacking any of these genes are termed “negative.” The collection of taxa is exposed to successive rounds of loss of gene function, converting a positive taxon into a negative taxon, chromosome rearrangement, in which randomly chosen single genes in a positive taxon may move independently to random chromosomal locations, and horizontal transfer, by which a negative taxon is converted to a positive taxon; the probability of transfer was inversely related to the distance separating all of genes required for the function (Figure 4). An effectively infinite pool of negative taxa was available as recipients of the horizontal transferred genes. The arrangement of genes within the newly created positive taxon is identical to that of the donor taxon. This process is repeated and the average distance separating the genes in question is determined. Clustering results, and the process is allowed to continue until the genes in question are separated by <1 min, on average, among the collection of positive taxa.

To investigate the properties of the Selfish Operon model, the computer model allows variation of the rates at which gene functions are lost, the rate of chromosomal rearrangement, the probability of horizontal

transfer, and the number of genes required for the function. The logic of the computer model reveals that genes must be horizontally transferred to avoid eventual loss [The purpose of the computer model is to examine the properties of the Selfish Operon model, and not to determine its veracity by rigorous simulation. Such a test would require estimates of parameters (*e.g.*, the population sizes and rates of horizontal transfer among species 2 billion years ago) that cannot be obtained.] Without opportunity for horizontal transfer, all taxa ultimately become negative. Since the probability of horizontal transfer is inversely related to the distance separating the loci (Figure 4), successive rounds of horizontal transfer allow more tightly clustered genes to dominate the collection of virtual taxa. The model was employed to determine how the average rate of gene clustering was influenced by the number of genes to be clustered, the rates of deletions and rearrangement, and the probability of horizontal transfer.

Two caveats should be noted. First, the process of the loss of gene function from any one taxon has been reduced to a single step; that is, a single event represents the sum of the initial mutation at one gene, its fixation in the population, and subsequent loss of all other genes in that pathway. Second, during any one simulation (that is, one run of the computer model as depicted in Figure 4), the loss of gene function was not allowed if very few positive taxa remained. This stipulation prevented all of the taxa from becoming negative (extinction of the function), leading to an unsuccessful termination of the simulation without gene clustering. In series of simulations without this condition, clustering still occurred, but greater numbers of simulations were required to obtain the same data.

*Gene clustering occurs in a three-step process:* The progress of a single simulation reflecting the clustering of three genes is shown in Figure 5. This simulation is typical of all simulations using this model and reflects three distinct phases. The initial shuffling phase is characterized by little net change in the average distance between loci in the collection of taxa. Since no gene clusters have arisen at this stage, there is little or no horizontal transfer of genes between taxa. However, as soon as a single taxon obtains a gene cluster by random gene shuffling, that cluster rapidly populates negative genomes. In concert with the inevitable loss of genes from all positive taxa, the average distance between genes steadily decreases among the collection of taxa. This period of decrease is designated the sweep phase. The sweep phase depicted in Figure 5 was initiated by a cluster of three genes positioned within an 8-min chromosome fragment. The rate of the sweep is dependent on the size of the fragment containing the gene cluster; tight clusters have a greater probability of transfer and sweep more quickly. Following the sweep phase, periodic sweeps by more tightly clustered genes occur; this is termed the cluster phase. The number of genera-

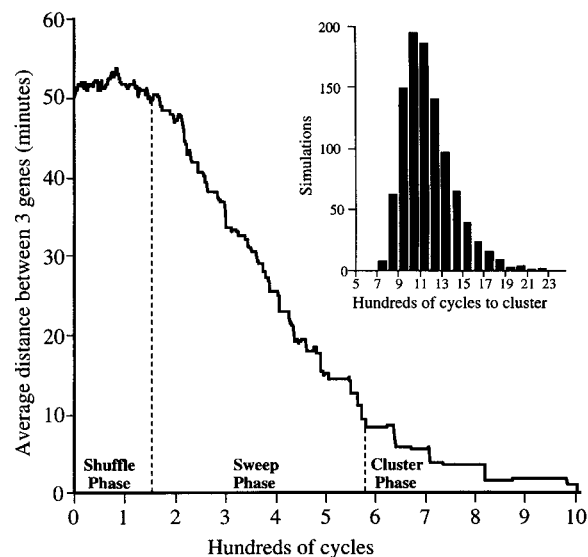


FIGURE 5.—The clustering of three gene following the model illustrated in Figure 4. The simulation was terminated at cycle 1005, when the average distance between the three loci was 0.3 min. The shuffle phase, sweep phase, and cluster phase are described in the text. For the simulation described here  $P_{loss} = 0.001$ ,  $P_{move} = 0.01$ , and maximal  $P_{transfer} = 0.01$  (see Figure 4). Inset histogram shows the times to clustering for three genes for 1000 simulations. Genes are considered “clustered” when the final average distance between alleles  $< 1$  min among the collection of positive species.

tions required for gene clustering can vary (inset in Figure 5); differences in the length of the cluster phase imparts much of this variance (data not shown).

The cluster phase is characterized by the rapid, successive transfer of newly created gene clusters to naïve genomes. The great length of the cluster phase (see Figure 5) is likely to be an indirect result of the simulation methods. As discussed above, the Selfish Operon model predicts that loosely clustered genes will aggregate rapidly once they are transferred to naïve genomes; this is due to the rapid removal of the nonselected, intervening DNA. When the deletion of intervening DNA is included in the simulation, the cluster phase is essentially eliminated. The overall behavior of the shuffle and sweep phases are not significantly altered.

*The time to gene clustering is a linear function of the number of genes:* One might expect that the time required for gene clustering would increase exponentially as more genes are required for the cluster. If so, the model would predict that clusters containing large numbers of genes would rarely be observed in extant genomes. However, simulations show that the time required for gene clustering is linearly related to the number of genes required for the function ( $R^2 = 0.997$ , Figure 6). The linear increase in the time to gene clustering reflects the time required to form the initial cluster, that is the shuffle phase. The length of the shuffle phase increases linearly as more genes are required to be physically proximal. However, the rate of the sweep phase

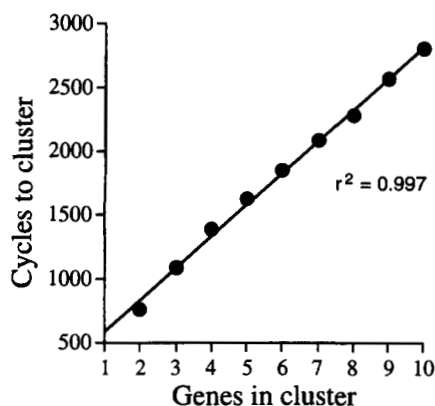


FIGURE 6.—The average time to clustering is plotted as a function of the number of genes in the cluster. Each point represents the average of 1000 simulations; parameters for the simulations were  $P_{loss} = 0.001$ ,  $P_{move} = 0.01$ , and maximal  $P_{transfer} = 0.01$  (see Figure 4). The regression line  $Y = 248.6X + 333$  has a correlation coefficient of 0.997.

is dependent on the size of the fragment containing the initial cluster, not on the number of genes in the cluster. Once the sweep phase is initiated, clusters of similar lengths sweep at similar rates. Therefore, this simulation predicts that naturally occurring clusters may harbor large numbers of genes. Inspection of the *E. coli* and *S. typhimurium* genetic maps reveal many operons of substantial size. The *S. typhimurium* *rfb* and *cob* operons, discussed above, contain 15 and 20 genes, respectively.

#### PREDICTIONS OF THE SELFISH OPERON MODEL

##### Genes for essential processes should not cluster:

The Selfish Operon model predicts that genes providing for essential processes should not be found in clusters. Mutant cells without an essential gene function could not survive in any environment, precluding the process of loss and acquisition posited by the Selfish Operon model. In contrast, the Coregulation model predicts that essential genes whose coregulation is most critical are those most likely to be found in operons. Surveys of bacterial chromosomes generally support the predictions of the Selfish Operon model. In general, genes providing for essential processes are not found in clusters. For example, the genes for NAD biosynthesis and cycling are found in  $\geq 13$  distinct loci in the *S. typhimurium* chromosome (SANDERSON *et al.* 1995); genes for DNA synthesis comprise  $\geq 11$  distinct loci in *E. coli* (BACHMANN 1990). A few exceptional cases of clustered essential genes, *e.g.*, the clusters of genes encoding ribosomal proteins, are discussed in detail below. The Selfish Operon model predicts that strongly clustered genes should be those that encode proteins for conditionally dispensable biosynthetic or degradative pathways for which selection can be relaxed for long periods of time.

##### Genes for nonessential functions are clustered: We

predict that many operons have formed during the diversification of bacterial species and have been present in host genomes for long periods of time. For example, many selfish operons may encode degradative pathways. Generally the encoded enzymes degrade a particular substance to produce a metabolite that easily enters central metabolism; the *S. typhimurium pdu* operon discussed above is one such operon. Genes for catabolic functions in *Pseudomonas aeruginosa* and *P. putida* are also clustered; these clusters are commonly located on plasmids (HOLLOWAY *et al.* 1990). HOLLOWAY and MORGAN (1986) suggested that the chromosomes of these species have been augmented by the introgression of such plasmids bearing clusters of useful genes. In *E. coli* (BACHMANN 1990), gene clusters encode functions for the degradation of lactose (*lacIZYA*), *n*-acetyl-glucosamine (*nagABE*), galactose (*galKTE*), proline (*putAP*), cellibiose (*celFDCB*), galactitol (*gatCAD*), glycerol-phosphate (*glpQTACB*),  $\gamma$ -amino-butyrate (*gabCPDT*), sorbitol (*srlABDR*), fucose (*fucDAPIKR*),  $\beta$ -galactosides (*ebgRACB*), hexuronates (*uxaAC*), xylose (*xyLABRF*), mannitol (*mitDAC*), galactonate (*dgoADKTR*),  $\beta$ -glucosides (*bglBFG*), ribose (*rbsABCDRK*), rhamnose (*rhaDABR*), melibiose (*melRAB*), and mannonate (*uxuRBA*). While these functions are useful, the failure to degrade any of these substances is counterselected only when that substance is present in high concentration.

Another class of transiently selected operons are those conferring transport functions. In *E. coli*, such operons include functions for the transport of ferri-chromes (*fhuABCD*), potassium (*kdpDCBA*), gluconate (*glnQPH*), oligopeptides (*oppABCD*), cobalamin (*btuCED*), mannose (*manXYZ*), arabinose (*araFGH*), methylgalactosides (*mglBACD*), histidine (*hisJPM*), proline (*proUVW*), glycerol-3-phosphate (*upgCEAB*), leucine, isoleucine, and valine (*livGMHKLJ*), hexose phosphates (*uhpTCBA*), and phosphates (*pstABCS*). Such transport functions may be considered selfish in that they would augment degradative or biosynthetic operons in recipient taxa.

Biosynthetic genes are only under selection when the end product of the encoded pathway is missing from the environment. In *E. coli*, such operons under non-continuous selection encode proteins for the synthesis of threonine (*thrABC*), betaine (*betABT*), tryptophan (*trpABCDE*), histidine (*hisGDCBHAFIE*), menaquinone (*menACEBD*), isoleucine and valine (*ilvGMEDAYC*), and thiamine (*thiABC*). The products of these operons are typically compounds available in the growth environment, or compounds not absolutely required for cell growth.

##### Recently introduced Selfish Operons can be detected:

If the transfer of the *cob* operon is typical of bacterial operons under weak selection, then analysis of bacterial chromosomes should reveal many operons that are of recent foreign ancestry. If genes are introduced into a chromosome from a donor with a significantly different

TABLE 4  
Atypical gene clusters in a 1.1-MB region of the *E. coli* chromosome

Group <sup>a</sup>	Gene <sup>b</sup>	Position <sup>c</sup>	Percent G + C	Group	Gene	Position	Percent G + C
A	<i>tdcC</i>	44440	48.8	K	<i>f242a</i>	444624	41.8
A	<i>tdcB</i>	45793	46.4	K	<i>f274a</i>	445720	41.0
A	<i>tdcA</i>	46881	42.2	L	<i>o173</i>	501406	46.2
A	<i>tdcR</i>	48008	30.7	L	<i>o283</i>	501924	42.9
A	<i>o186</i>	48587	36.0	M	<i>yiaA</i>	507650	41.3
A	<i>o395</i>	49169	31.8	M	<i>yiaB</i>	508133	41.5
B	<i>o194</i>	68178	45.1	N	<i>o155</i>	524470	45.3
B	<i>o232</i>	68842	44.3	N	<i>o157a</i>	525055	44.1
B	<i>o863</i>	69491	47.5	O	<i>yibB</i>	573551	36.9
B	<i>o429</i>	72093	47.9	O	<i>rfaD</i>	574726	50.9
C	<i>o224</i>	142857	41.6	O	<i>rfaF</i>	575668	53.7
C	<i>o793</i>	143552	45.5	O	<i>rfaC</i>	576718	51.0
C	<i>o159</i>	145930	45.4	O	<i>rfaL</i>	577687	36.2
D	<i>o104</i>	166232	47.3	O	<i>rfaK</i>	578978	31.0
D	<i>f90</i>	166602	42.9	O	<i>rfaZ</i>	580084	32.5
E	<i>o59</i>	193367	40.6	O	<i>rfaY</i>	581006	35.3
E	<i>f220</i>	193549	44.2	O	<i>rfaJ</i>	581722	33.8
F	<i>f139</i>	234256	46.0	O	<i>rfaI</i>	582778	36.2
F	<i>f489</i>	234667	49.0	O	<i>rfaB</i>	583797	39.4
F	<i>o271</i>	236326	42.4	O	<i>rfaS</i>	584920	26.8
G	<i>o392</i>	362600	34.0	O	<i>rfaP</i>	585892	43.4
G	<i>o138</i>	163775	29.7	O	<i>rfaG</i>	586682	44.8
H	<i>yhhH</i>	404151	32.8	O	<i>rfaQ</i>	587803	44.6
H	<i>yhhI</i>	405115	41.5	P	<i>o155</i>	678247	43.8
I	<i>f450</i>	411703	37.5	P	<i>o195</i>	678763	44.6
I	<i>f123</i>	413587	35.5	Q	<i>f126</i>	783120	34.9
I	<i>f409</i>	413963	34.7	Q	<i>f138</i>	783514	33.3
J	<i>o260</i>	432025	36.0	R	<i>o326</i>	841805	43.9
J	<i>IS50R</i>	432916	54.7	R	<i>o421</i>	842887	45.0
J	<i>slp</i>	434662	44.7	R	<i>f723</i>	844243	43.7
J	<i>yhiF</i>	435417	39.7	S	<i>o81</i>	859903	44.7
J	<i>f215</i>	435989	45.5	S	<i>o80</i>	860366	45.7
J	<i>f112</i>	436700	39.8	T	<i>o99</i>	893600	46.3
J	<i>f110</i>	437142	42.9	T	<i>o142</i>	893926	43.1
J	<i>o190</i>	437729	46.6	U	<i>o84</i>	1040956	28.6
J	<i>o175</i>	439100	33.0	U	<i>o235</i>	1041234	36.6

<sup>a</sup> All genes comprising a group are contiguous; no intervening genes have been eliminated.

<sup>b</sup> Gene designations were described in the corresponding GenBank entry.

<sup>c</sup> Position in contiguous sequence beginning with GenBank sequence U18997 (through base 367560) and continuing through sequences U00039 (through base 491891), L10328 (through base 721915), M87049 (through base 815802), L19201 (through base 916648), and U00006 (through base 1090099); the region corresponds to minutes 67–92 on the *E. coli* genetic map.

DNA composition, these genes can be identified by their unusual DNA sequence features, such as base composition, codon usage bias, and dinucleotide fingerprints. If the transfer were recent, then the DNA sequence would not have had time to ameliorate, that is to adjust their composition to resemble the surrounding DNA (OCHMAN and LAWRENCE 1996). These gene clusters may represent selfish clusters of recent introgression. The *cob*, *pdu*, *rfb*, *spa*, and *oad* operons discussed above (and listed in Table 3) provide some examples of introgressed operons in the *S. typhimurium* chromosome; each operon shows an unusual DNA composition.

A more systematic analysis of the *E. coli* chromosome reveals numerous gene clusters that appear atypical

(Table 4). As expected, these clusters are few in number since the majority of the bacterial chromosome is inherited vertically. To detect recently introgressed clusters, the foreign DNA must be distinguishable from genes native to the host chromosome. Using established criteria (OCHMAN and LAWRENCE 1996), we have analyzed 1,100,000 contiguous bases of the *E. coli* chromosome with respect to GC content, dinucleotide frequencies, and codon usage bias. Aberrant genes were detected as those with GC contents in the first and third codon positions  $\geq 10\%$  lower, or 8% higher, than the averages for these position (59 and 55% G+C, respectively) in genes of this organism, a low codon usage bias (CAI < 0.3 by the method of SHARP and LI 1987), and a chi-square for codon usage that is twofold higher than pre-

dicted for genes of a corresponding CAI (see OCHMAN and LAWRENCE 1996). Dinucleotide frequencies of aberrant genes were also significantly different from those of typical *E. coli* genes.

The 1.1 Mb contiguous sequence from *E. coli* contained 976 open reading frames (ORF); using conservative parameters, 109 ORFs were preliminarily identified as being atypical using the criteria described above. While some atypical genes were not found in clusters, clusters of 2–15 genes were identified among these sequences (Table 4). Some of these genes have been identified and assigned functions by previous workers. The *rfa* operon encodes proteins responsible for O-antigen formation. The *tdc* operon encodes genes for the transport and dehydration of threonine. However, many of the clusters with aberrant sequence characteristics contained genes with unknown functions. These data support the prediction that horizontally transferred genes are introduced in clusters.

**Selfish Operons are consistent with *E. coli* population biology:** If the process of gene clustering by horizontal transfer outlined by Selfish Operon model is occurring among extant bacterial species, then one would predict that bacterial chromosomes are in constant genomic flux. That is, genes are being continuously added by horizontal transfer and lost by mutation and genetic drift. The analysis of the 1.1 Mb fragment of the *E. coli* chromosome shows that a substantial number of genes have been recently introduced into the *E. coli* genome. If one assumes that bacterial chromosomes are not growing continually larger in size, then a similar amount of DNA must have been recently lost from the *E. coli* chromosome. This genomic flux depicts a dynamic bacterial chromosome in a constant state of change. This model is somewhat different from the historical expectation of a relatively unchanging genome.

The nearly identical genetic maps of the related bacteria *E. coli* and *S. typhimurium* have historically supported the model of an evolutionarily stable chromosome. Despite the presence of conjugative plasmids, transducing bacteriophages, transposable elements, and molecular mechanisms for chromosome rearrangement, the genomes of these species are remarkably similar with respect to DNA composition (~51% G+C), the order of shared genes (BACHMANN 1990; SANDERSON *et al.* 1995) and average size [~4.8 Mb; much of the intraspecific variation in genome size has been attributed to the introgression of episomes and to horizontal transfer (HARSONO *et al.* 1993; BERGTHORSSON and OCHMAN 1995)]. The remarkable similarity of the genomes of the two taxa was taken as evidence that bacterial chromosomes are not in a state of rapid change. Moreover, natural populations of *E. coli* show substantial linkage disequilibrium based on enzyme electrophoretotypes; these data support a model of a clonal bacterial population with frequent periodic selection maintaining dominant clones (SELANDER and LEVIN 1980;

LEVIN 1981; ACHTMAN *et al.* 1981; OCHMAN and SELANDER 1984; WHITTAM *et al.* 1984). Occasional intraspecific recombination (MILKMAN and MCKANE BRIDGES 1990, 1993; BOYD *et al.* 1994; GUTTMAN and DYKHUIZEN 1994) is apparently insufficient to disrupt the linkage disequilibrium. These data serve to reinforce the inference of slow chromosome change reflecting the clonal nature of prokaryotic reproduction. The observations that heavily transcribed genes (BREWER 1990) and the  $\chi$  recombination stimulation sites (BURLAND *et al.* 1993) are preferentially oriented *vis-à-vis* the origin of replication have been interpreted as requiring this evolutionarily stable state to evolve.

The dynamic introduction of novel DNA into bacterial chromosomes envisioned by the Selfish Operon model is not inconsistent with the data suggesting a static, slowly evolving bacterial chromosome. The static natures of the *E. coli* and *S. typhimurium* chromosomes were inferred by comparing the map positions of a small number of genetic loci shared between these taxa; genes unique to one taxon could be interpreted as being either uncharacterized in or deleted from the other taxon. However, horizontal transfer events can also explain these species-specific functions. Introduction of novel genetic material neither alters the order of the surrounding genes nor disrupts linkage disequilibrium. The enzymes that suggested strong linkage disequilibrium, when assayed by multilocus enzyme electrophoresis, are typically encoded by essential genes and are involved in central metabolic pathways; such pathways would be generally unaffected by the introgression of foreign DNA. Loci subject to diversifying selection, like the *gnd* gene, show unusually high levels of variation; this variation does not affect the chromosomal disequilibrium. We conclude that the Selfish Operon model is compatible with the previous observations that suggested evolutionarily stable chromosomes.

#### INCONSISTENCIES WITH THE SELFISH OPERON MODEL

**Certain essential genes are clustered:** The Selfish Operon model predicts that essential genes will be unclustered. While this is generally true, there are large gene clusters that are universally agreed upon as being indispensable to all bacterial cells. Notable examples include the clusters of genes encoding ribosomal proteins (*rpl*, *rps*) and the cluster of genes encoding the  $F_1F_0$  ATPase (*atp*). The clustering of these genes can be explained in two ways. First, these gene clusters may have formed as the result of very different selective processes than those that formed most gene clusters; it is likely that these clusters formed prior to the divergence of all known life. Substantial portions of the *E. coli* and *B. subtilis* eubacterial ribosomal protein gene clusters are nearly identical to the ribosomal protein gene clusters of the archaeobacteria *Halobacterium marismortui* and

*Methanococcus vannielii* (WITTMAN-LIEBOLD *et al.* 1990); the gene order within the *atp* operons of *E. coli* and *H. influenza* are also nearly identical (FLEISCHMANN *et al.* 1995). These data support the hypothesis that these operons were assembled before the divergence of all known life. Speculation on how natural selection may have operated on primordial cells is less than robust; many models may be applied to the formation of gene clusters under primordial conditions.

However, the *atp* genes and genes encoding ribosomal proteins share a notable feature: these operons encode groups of proteins that must interact physically. Therefore, it is likely that particular alleles of Atp proteins may be coadapted to work well together; in a similar fashion, alleles of ribosomal proteins may be coadapted to work together. Hence, one may view the *rps*, *rpl*, and *atp* operons as coadapted gene complexes; other coadapted genes complexes include the genes for bacteriophage head proteins. As discussed above (*Fisher model revisited*), the Selfish Operon model predicts that successful horizontal transfer of a coadapted gene complexes requires that all genes are transferred together. In this fashion, native, unclustered *rps* genes may be superseded by a horizontally transferred, coadapted *rps* gene complex that is favored by natural selection; such selection may include greater translational efficiency or the resistance to antibiotics. Since this process does not entail the loss of the native *rps* gene complex before the horizontal transfer event, Selfish Coadapted Operons may include essential genes.

**Some nonessential genes with related functions are not clustered:** While some loci are clustered in the *S. typhimurium* chromosome, *e.g.*, the *his*, *ilv*, *leu*, and *trp* amino acid biosynthetic genes (Table 1), genes encoding other biosynthetic pathways are dispersed, *e.g.*, the *cys* and *met* amino acid biosynthetic genes (Table 2). We infer that the *cys* and *met* loci have never been lost and reacquired by ancestors of the *Salmonella* lineage and conclude that cysteine and methionine participate in cellular metabolism in a manner that is significantly different than other amino acids. Histidine, isoleucine, valine, leucine, and tryptophan participate only in protein synthesis. In contrast, cysteine and methionine participate in additional cellular metabolism. Cysteine serves as the source of reduced sulfur for the methionine, glutathione, thiamine, and biotin biosynthetic pathways; methionine is incorporated into S-adenosyl-methionine, an important methyl-group donor. Therefore, transient starvation for these compounds may be strongly counterselected. The additional roles of cysteine and methionine in cellular metabolism provide sufficient selection to prevent loss and reacquisition of their biosynthetic genes. Alternatively, these functions may be essential in nature if these compounds cannot be readily obtained from the environment.

**Genes are clustered in certain species, but not in**

**others:** The degree of gene clustering among some loci may reflect historical or lineage-specific processes. For example, while the *pur* genes are scattered in the *S. typhimurium* chromosome (SANDERSON *et al.* 1995), they are found in single 11-gene cluster in the *Bacillus subtilis* chromosome (EBBOLE and ZALKIN 1987; SAXILD and NYGAARD 1988). The degree to which *trp* genes are clustered varies widely among bacterial lineages; for example, the six *trp* genes form a single operon in *E. coli* but are found in three or four clusters among various species of *Pseudomonas* (CRAWFORD 1989; CRAWFORD and MILKMAN 1991). One may consider it unlikely that purine synthesis was under weak selection in the *B. subtilis* lineage, leading to loss and reacquisition of the *pur* genes as a single operon. However, the natural environments of some bacteria may result in weak selection for some processes deemed indispensable for other species. For example, in natural populations *Salmonella* mutants have been found which are deficient in pathways that are intuitively indispensable; STOKES and BAYNE (1958) identified purine-requiring, thiamine-requiring, and nicotinate-requiring strains of *S. typhimurium*, *S. paratyphi*, and *S. abortusovis*. As a result, some lineages may bear clusters of genes not found in clusters in other lineages. As detailed above, other genes clusters may reflect selection for sets of coadapted products. The cluster of 27 tRNA genes in *Staphylococcus aureus* (GREEN and VOLD 1993) may encode a suite of tRNAs that provide efficient translation of genes bearing a particular codon usage bias.

**Some operons contain apparently unrelated genes:** Several operons have been characterized that include genes of apparently unrelated functions. For example, the *rplKAJL* genes of the cyanobacterium *Synechocystis* PCC6803 encode four ribosomal proteins. These genes are cotranscribed with the *aroC* gene, which encodes chorismate synthase (SCHMIDT *et al.* 1993). The *rplKAJL* genes of *E. coli* are cotranscribed with the *rpoBC* genes, which encode subunits of RNA polymerase (BACHMANN 1990). It is possible that the functions of these genes are related in an as-yet-undetermined fashion. Alternatively, the clustering of these genes may have been purely accidental. One would expect that small chromosomal deletions occasional fuse transcripts of functionally unrelated genes. Such arrangements would persist if the new operons were not counterselected. One would expect the genes of these unusual operons to be amenable to coordinate expression, that is, the conditions that repress or derepress the new cluster are not in conflict with the conditions that should repress or derepress some of the constituent genes. Alternatively, these operons could be constitutively expressed. This would appear to be true for the *rplKAJL* clusters of *E. coli* and *Synechocystis* PCC6903.

**Conclusions:** The Selfish Operon model provides a mechanism for the clustering of genes contributing to a single function; the model does not require that the

gene cluster provide any selective benefit to the host organism. While this model is consistent with a majority of gene clusters observed in extant bacterial chromosomes, there are cases which are more difficult to understand. As GOLDSCHMIDT (1951) noted, "a theory trying to unify a vast and difficult field with innumerable details is certainly nothing static; it is a fleeting moment in an eternal flux." Hence, the degree to which horizontal transfer has contributed to the overall level of gene clustering in bacterial chromosomes cannot be stated precisely.

We thank T. BOBIK, D. DYKHUIZEN, T. GALITSKI, E. KOFOID, H. OCHMAN, E. PRESLEY and P. SHARP for helpful, critical, and enlightening discussions and A. CAMPBELL, R. MILKMAN, and W.-H. LI for helpful comments on the manuscript. This work was supported by grants GM-15868 (J.G.L.) and GM-34804 (J.R.R.) from the National Institutes of Health.

#### LITERATURE CITED

- ACHTMAN, M., A. MERCER, B. KUSECEK, A. POHL, M. HEUZENROEDER *et al.*, 1983 Six widespread bacterial clones among *E. coli* K1 isolates. *Infect. Immun.* **39**: 315–335.
- AMES, B. N., and R. G. MARTIN, 1964 Biochemical aspects of genetics: the operon. *Annu. Rev. Biochem.* **33**: 235–258.
- BACHMANN, B., 1990 Linkage map of *Escherichia coli* K-12, Edition 8. *Microbiol. Rev.* **54**: 130–197.
- BANNISTER, J. V., and M. W. PARKER, 1985 The presence of a copper/zinc superoxide dismutase in the bacterium *Photobacterium leignathi*: a likely case of gene transfer from eukaryotes to prokaryotes. *Proc. Natl. Acad. Sci. USA* **82**: 149–152.
- BARRATT, W. R., D. NEUMEYER, D. D. PERKINS and L. GARNJOBST, 1954 Map construction in *Neurospora crassa*. *Adv. Genet.* **6**: 1–93.
- BEADLE, G. W., and E. L. TATUM, 1941 Genetic control of biochemical reactions in *Neurospora*. *Proc. Natl. Acad. Sci. USA* **27**: 499–596.
- BECKWITH, J. R., A. R. PARDEE, R. AUSTRIAN and F. JACOB, 1962 Coordination of synthesis of the enzymes in the pyrimidine pathway of *E. coli*. *J. Mol. Biol.* **5**: 618–634.
- BELFAZIA, J., C. PARSOT, A. MARTEL, C. BOUTHER DE LA TOUR, D. MARGARITA *et al.*, 1986 Evolution in biosynthetic pathways: two enzymes catalyzing consecutive steps in methionine biosynthesis originate from a common ancestor and possess a similar regulatory region. *Proc. Natl. Acad. Sci. USA* **83**: 867–871.
- BERGTHORSSON, U., and H. OCHMAN, 1995 Heterogeneity of genome size among natural isolates of *Escherichia coli*. *J. Bacteriol.* **177**: 5784–5789.
- BLUNDELL, M., E. CRAIG and D. KENNEL, 1972 Decay rates of different mRNA in *E. coli* and models of decay. *Nature New Biol.* **238**: 46–49.
- BOBIK, T. A., M. AILION and J. R. ROTH, 1992 A single regulatory gene integrates control of vitamin B<sub>12</sub> synthesis and propanediol degradation. *J. Bacteriol.* **174**: 2253–2266.
- BODMER, W. F., and P. A. PARSONS, 1962 Linkage and recombination in evolution. *Adv. Genet.* **11**: 1–100.
- BOTSTEIN, D., 1980 A theory of modular evolution for bacteriophages. *Ann. NY Acad. Sci.* **354**: 484–491.
- BOYD, E. F., K. NELSON, F. S. WANG, T. S. WHITTAM and R. K. SELANDER, 1994 Molecular genetic basis of allelic polymorphism in malate dehydrogenase (*mdh*) in natural populations of *Escherichia coli* and *Salmonella enterica*. *Proc. Natl. Acad. Sci. USA* **91**: 1280–1284.
- BREWER, B. J., 1990 Replication and the transcriptional organization of the *Escherichia coli* chromosome, pp. 61–83 in *The Bacterial Chromosome*, edited by K. DRLICA and M. RILEY. American Society for Microbiology, Washington, DC.
- BRIDGES, C. B., and K. S. BREHME, 1944 The mutants of *Drosophila melanogaster*. *Carnegie Inst. Washington Publ.* **522**.
- BRISSON-NOEL, A., M. ARTHUR and P. COURVALIN, 1988 Evidence for natural gene transfer from gram-positive cocci to *Escherichia coli*. *J. Bacteriol.* **170**: 1739–1745.
- BRÜCKNER, R., and H. MATZURA, 1981 In vivo synthesis of a polycistronic messenger RNA for the ribosomal proteins L11, L1, L10, and L7/12 in *Escherichia coli*. *Mol. Gen. Genet.* **183**: 277–282.
- BUEHNER, M., G. C. FORD, D. MORAS, K. W. OLSEN and M. G. ROSSMAN, 1973 D-Glyceraldehyde-3-phosphate dehydrogenase: three-dimensional structure and evolutionary significance. *Proc. Natl. Acad. Sci. USA* **70**: 3052–3054.
- BURLAND, V., F. PLUNKETT, D. DANIELS and F. R. BLATTNER, 1993 DNA sequence and analysis of 136 kilobases of the *Escherichia coli* genome: organizational symmetry around the origin of replication. *Genomics* **16**: 551–561.
- BUVINGER, W. E., K. A. LAMPEL, R. J. BOJANOWSKI and M. RILEY, 1984 Location and analysis of nucleotide sequences at one end of a putative *lac* transposon in the *Escherichia coli* chromosome. *J. Bacteriol.* **159**: 618–623.
- CAMPBELL, A., and D. BOTSTEIN, 1983 Evolution of lambdoid phages, pp. 365–380 in *Lambda II*, edited by R. W. HENDRIX, J. W. ROBERTS, F. W. STAHL, and R. A. WEISBERG. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- CASJENS, S., 1974 Bacteriophage lambda *FII* gene protein: role in head assembly. *J. Mol. Biol.* **90**: 1–23.
- CASJENS, S., and R. HENDRIX, 1988 Control mechanisms in dsDNA bacteriophage assembly, pp. 15–91 in *The Bacteriophages*, edited by R. CALENDER. Plenum Publishing, New York.
- CASJENS, S., G. HATFULL and R. HENDRIX, 1992 Evolution of dsDNA tailed-bacteriophage genomes. *Virology* **3**: 383–397.
- CRAWFORD, I. P., 1989 Evolution of a biosynthetic pathway: the tryptophan paradigm. *Annu. Rev. Microbiol.* **43**: 567–600.
- CRAWFORD, I. P., and R. MILKMAN, 1991 Orthologous and paralogous divergence, reticulate evolution, and lateral gene transfer in bacterial *trp* genes, pp. 77–95 in *Evolution at the Molecular Level*, edited by R. K. SELANDER, A. G. CLARK, and T. S. WHITTAM. Sinauer Associates, Sunderland, MA.
- DEMEREK, M., 1960 Frequency of deletions among spontaneous and induced mutations in *Salmonella*. *Proc. Natl. Acad. Sci. USA* **46**: 1075–1079.
- DEMEREK, M., and Z. E. DEMEREK, 1956 Analysis of linkage relationships in *Salmonella* by transduction techniques. *Brookhaven Symp. Biol.* **8**: 75–84.
- DEMEREK, M., and Z. E. HARTMAN, 1956 Tryptophan mutants in *Salmonella typhimurium*. *Carnegie Inst. Washington Publ.* **612**: 5–33.
- DEMEREK, M., and P. HARTMAN, 1959 Complex loci in microorganisms. *Annu. Rev. Microbiol.* **13**: 377–406.
- DEMEREK, M., I. BLOOMSTRAND and Z. E. DEMEREK, 1955 Evidence of complex loci in *Salmonella*. *Proc. Natl. Acad. Sci. USA* **41**: 359–364.
- DEMEREK, M., E. L. LAHR, E. BALBINDER, T. MIYAKE, C. MACK *et al.*, 1959 Bacterial genetics. *Carnegie Inst. Washington Year Book* **58**: 433–439.
- DESJARDINS, P., B. PICARD, B. KALTENBÖCH, J. ELION and E. DENAMUR, 1995 Sex in *Escherichia coli* does not disrupt the clonal structure of the population: evidence from random amplified polymorphic DNA and restriction-fragment-length polymorphism. *J. Mol. Evol.* **41**: 440–448.
- DOOLITTLE, R. F., D. F. FENG, K. L. ANDERSON and M. R. ALBERRO, 1990 A naturally occurring horizontal gene transfer from a eukaryote to a prokaryote. *J. Mol. Evol.* **31**: 383–388.
- DUNN, L. C., and E. CASPARI, 1945 A case of neighboring loci with similar effects. *Genetics* **30**: 543–568.
- DUNN, L. C., 1954 The study of complex loci. *Caryologia* **6**: 155–166.
- EBBOLE, D. J., and H. ZALKIN, 1987 Cloning and characterization of a 12 gene cluster from *Bacillus subtilis* encoding nine enzymes for *de novo* purine nucleotide synthesis. *J. Biol. Chem.* **262**: 8274–8287.
- FANI, R., P. LIÒ, I. CHIARELLI and M. BAZZICALUPO, 1994 The evolution of the histidine biosynthetic genes in prokaryotes: a common ancestor for the *hisA* and *hisF* genes. *J. Mol. Evol.* **38**: 489–495.
- FAZZIO, T. G., and J. R. ROTH, 1996 Evidence that the CysG protein catalyzes the first reaction specific to B<sub>12</sub> synthesis in *Salmonella typhimurium*: insertion of cobalt. *J. Bacteriol.* (in press).
- FINCHAM, J. R. S., and J. A. PATEMAN, 1957 Formation of an enzyme

- through complementary action of mutant "alleles" in separate nuclei in a heterocaryon. *Nature* **179**: 741–742.
- FISHER, R. A. 1930 *The Genetical Theory of Natural Selection*. Oxford University Press, Oxford.
- FLEISCHMANN, R. D., M. D. ADAMS, O. WHITE, R. A. CLAYTON, E. F. KIRKNESS *et al.*, 1995 Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**: 496–512.
- GLANVILLE, E. V., and M. DEMEREC, 1960 Threonine, isoleucine, and isoleucine-valine mutants of *Salmonella typhimurium*. *Genetics* **45**: 1359–1374.
- GOLDSCHMIDT, R., 1951 The theory of the gene. Cold Spring Harbor Symp. Quant. Biol. **16**: 1–11.
- GORINI, L., W. GUNDERSEN and M. BURGER, 1962 Genetics of regulation of enzyme synthesis in the arginine biosynthetic pathway of *Escherichia coli*. Cold Spring Harbor Symp. Quant. Biol. **26**: 173–182.
- GREEN, C. J., and B. S. VOLD, 1993 *Staphylococcus aureus* has clustered tRNA genes. *J. Bacteriol.* **175**: 5091–5096.
- GROISMAN, E. A., and H. OCHMAN, 1993 Cognate gene clusters govern invasion of host epithelial cells by *Salmonella typhimurium* and *Shigella flexneri*. *EMBO J.* **12**: 3779–3787.
- GRÜNBERG, H., 1935 Gene doublets as evidence for adjacent small duplications in *Drosophila*. *Nature* **140**: 932.
- GUTTMAN, D. S., and D. E. DYKHUIZEN, 1994 Detecting selective sweeps in naturally occurring *Escherichia coli*. *Genetics* **138**: 993–1003.
- HARDY, S. J. S., 1975 The stoichiometry of the ribosomal proteins of *Escherichia coli*. *Mol. Gen. Genet.* **140**: 253–274.
- HARSONO, K. D., C. W. CASPAR and J. B. LUCHANSKY, 1993 Comparison and genomic sizing of *Escherichia coli* O157:H7 isolates by pulsed-field gel electrophoresis. *Appl. Environ. Microbiol.* **59**: 3141–3144.
- HARTL, D. L., E. R. LOZOVSKAYA and J. G. LAWRENCE, 1992 Nonautonomous transposable elements in prokaryotes and eukaryotes. *Genetica* **86**: 47–53.
- HARTMAN, P. E., 1956 Linked loci in the control of consecutive steps in the primary pathway of histidine synthesis in *Salmonella typhimurium*. *Carnegie Inst. Washington Publ.* **612**: 35–62.
- Hartman, P. E., J. C. Loper and D. Šerman, 1960 Fine structure mapping by complete transduction between histidine-requiring *Salmonella* mutants. *J. Gen. Microbiol.* **22**: 323–353.
- HILDEBRANDT, V., M. RAMEZANI-RAD, U. SWIDA, P. WREDE, S. GRZESIEK *et al.*, 1989 Genetic transfer of the pigment bacteriorhodopsin into the eukaryote *Schizosaccharomyces pombe*. *FEBS Lett.* **243**: 137–140.
- HOLLOWAY, B. W., and A. F. MORGAN, 1986 Genome organization in *Pseudomonas*. *Annu. Rev. Microbiol.* **40**: 79–105.
- HOLLOWAY, B. W., S. DHARMSHITI, V. KRISHNAPPILLI, A. MORGAN, V. OBEYSEKERE *et al.*, 1990 Patterns of gene linkages in *Pseudomonas* species, pp. 97–105 in *The Bacterial Chromosome*, edited by K. DRLICA and M. RILEY. American Society for Microbiology, Washington, DC.
- HOLM, L., and C. SANDER, 1994 The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res.* **22**: 3600–3609.
- HOROWITZ, N. H., 1945 On the evolution of biochemical synthesis. *Proc. Natl. Acad. Sci. USA* **31**: 153–157.
- HOROWITZ, N. H., 1965 The evolution of biochemical syntheses—retrospect and prospect. Pp 15–23 in *Evolving genes and proteins*, edited by V. BRYSON and H. J. VOGEL. Academic Press, New York.
- HOROWITZ, N. H., and U. LEUPOLD, 1961 Some recent studies bearing on the one gene-one enzyme hypothesis. Cold Spring Harbor Symp. Quant. Biol. **16**: 75–74.
- HOWARTH, S., 1958 Suppressor mutations in some cystine-requiring mutants of *Salmonella typhimurium*. *Genetics* **43**: 404–418.
- HURST, G. D. D., L. D. HURST and M. E. N. MAJARUS, 1992 Selfish genes move sideways. *Nature* **356**: 659–660.
- JACOB, F., D. PERRIN, C. SANCHEZ and J. MONOD, 1960 L'opéron: groupe de gènes à expression coordonnée par un opérateur. *C. R. Acad. Sci.* **250**: 1727–1729.
- JACOB, F., and J. MONOD, 1962 On the regulation of gene activity. Cold Spring Harbor Symp. Quant. Biol. **26**: 193–211.
- JETER, R. M., B. M. OLIVERA and J. R. ROTH, 1984 *Salmonella typhimurium* synthesizes cobalamin (vitamin B<sub>12</sub>) *de novo* under anaerobic growth conditions. *J. Bacteriol.* **159**: 206–213.
- JONES, B. K., B. R. MONKS, S. A. LIEBHABER and N. E. COOKE, 1995 The human growth hormone gene is regulated by a multicomponent locus control region. *Mol. Cell. Biol.* **15**: 7010–7021.
- KÄFER, E., 1958 An eight-chromosome map of *Aspergillus nidulans*. *Adv. Genet.* **9**: 105–145.
- KEPES, A., 1967 Sequential transcription and translation in the lactose operon of *Escherichia coli*. *Biochim. Biophys. Acta* **138**: 107–123.
- KIDWELL, M., 1993 Lateral transfer in natural populations of eukaryotes. *Annu. Rev. Genet.* **27**: 235–256.
- KIMURA, M., 1983 *The Neutral Allele Theory of Molecular Evolution*. Cambridge University Press, Cambridge.
- KLENA, J. D., E. PRADEL and C. A. SCHNAITMAN, 1993 The *rfaS* gene, which is involved in production of a rough form of lipopolysaccharide core in *Escherichia coli* K-12, is not present in the *rfa* cluster of *Salmonella typhimurium* LT2. *J. Bacteriol.* **175**: 1524–1527.
- KOMAI, T., 1950 Semi-allelic genes. *Am. Natur.* **84**: 381–392.
- LAWRENCE, J. G., and J. R. ROTH, 1995 The cobalamin (coenzyme B<sub>12</sub>) biosynthetic genes of *Escherichia coli*. *J. Bacteriol.* **177**: 6371–6380.
- LAWRENCE, J. G., and J. R. ROTH, 1996 Evolution of coenzyme B<sub>12</sub> synthesis among enteric bacteria: evidence for loss and reacquisition of a multigene complex. *Genetics* **142**: 11–24.
- LEDERBERG, E., 1952 Allelic relationships and reverse mutation in *Escherichia coli*. *Genetics* **37**: 69–483.
- LEDERBERG, E. M. 1960 Genetic and functional aspects of galactose metabolism in *Escherichia coli* K-12, pp. 115–131 in *Microbial Genetics, Tenth Symposium of The Society for General Microbiology*, Vol. 113. Cambridge University Press, London.
- LEE, N., and E. ENGBERG, 1962 Dual effects of structural genes in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **48**: 335–348.
- LEWIS, E. B., 1947 Pseudoallelism in *Drosophila melanogaster*. *Genetics* **33**: 113.
- LEWIS, E. B., 1951 Pseudoallelism and gene evolution. Cold Spring Harbor Symp. Quant. Biol. **16**: 159–174.
- LEVIN, B., 1981 Periodic selection, infectious gene exchange, and the genetic structure of *E. coli* populations. *Genetics* **99**: 1–23.
- LIU, D., and P. R. REEVES, 1994 Presence of different O antigen forms in three isolates of one clone of *Escherichia coli*. *Genetics* **138**: 6–10.
- MANIATIS, T., E. F. FRITSCH, J. LAUER and R. M. LAWN, 1980 The molecular genetics of human hemoglobins. *Annu. Rev. Genet.* **14**: 145–178.
- MARGOLIN, P., J. VANDERMEULEN-COCITO and E. J. SIMRELL, 1959 The leucine locus of *Salmonella typhimurium*. *Annu. Report Biol. Lab.* **1959**: 41–43.
- MARTIN, R. G., M. A. BERBERICH, B. N. AMES, W. W. DAVIS, R. F. GOLDBERGER and J. YOURNO, 1971 Enzymes and intermediates of histidine biosynthesis in *Salmonella typhimurium*. *Methods Enzymol.* **17**: 3–44.
- MARTIN, J., and R. E. WEBSTER, 1975 The *in vitro* translation of a terminating signal by a single *Escherichia coli* ribosome. *J. Biol. Chem.* **250**: 8132–8139.
- MATSUI, K., K. SANO and E. OHTSUBO, 1986 Complete nucleotide and deduced amino acid sequences of the *Brevibacterium lactofermentum* tryptophan operon. *Nucleic Acids Res.* **14**: 10113–10114.
- MAYNARD SMITH, J., N. H. SMITH, M. O'ROURKE and B. G. SPRATT, 1993 How clonal are bacteria? *Proc. Natl. Acad. Sci. USA* **90**: 4384–4388.
- MÉDIGUE, C., T. ROUXEL, P. VIGIER, A. HÉNAUT and A. DANCHIN, 1991 Evidence of horizontal gene transfer in *Escherichia coli* speciation. *J. Mol. Biol.* **222**: 851–856.
- MILKMAN, R., and M. MCKANE BRIDGES, 1990 Molecular evolution of the *E. coli* chromosome. III. Clonal frames. *Genetics* **126**: 505–517.
- MILKMAN, R. and M. MCKANE BRIDGES, 1993 Molecular evolution of the *E. coli* chromosome. IV. Sequence comparisons. *Genetics* **133**: 455–468.
- MITCHELL, M. B., 1955 Aberrant recombination of pyridoxine mutants in *Neurospora crassa*. *Proc. Natl. Acad. Sci. USA* **41**: 215–220.
- MITCHELL, M. B., and H. K. MITCHELL, 1950 The selective advantage of an adenineless double mutant over one of the single mutants involved. *Proc. Natl. Acad. Sci. USA* **36**: 115–119.
- MIYAKE, T., and M. DEMEREC, 1960 Proline mutants of *Salmonella typhimurium*. *Genetics* **45**: 755–762.

- MOURANT, A. E., 1971 Transduction and skeletal evolution. *Nature* **231**: 466–467.
- NEWMAYER, D., 1957 Arginine synthesis in *Neurospora crassa*: genetic studies. *J. Gen. Microbiol.* **16**: 449–462.
- NOLAN, C., and E. MARGOLASH, 1968 Comparative aspects of primary structures of proteins. *Annu. Rev. Biochem.* **23**: 727–792.
- OCHMAN, H., and J. G. LAWRENCE, 1996 Phylogenetics and the amelioration of bacterial genomes, pp. 2627–2637 in *Escherichia coli and Salmonella: Cellular and Molecular Biology*, Ed. 2, edited by F. C. NEIDHARDT, R. CURTISS III, J. L. INGRAHAM, E. C. C. LIN, K. BROOKS LOW *et al.* ASM Press, Washington, DC.
- OCHMAN, H., and R. K. SELANDER, 1984 Evidence for clonal population structure in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **81**: 198–201.
- OPPENHEIM, D. S., and C. YANOFSKY, 1980 Translational coupling during expression of the tryptophan operon of *Escherichia coli*. *Genetics* **95**: 785–795.
- ORENGO, C. A., T. P. FLORES, W. R. TAYLOR and J. M. THORTON, 1993 Identification and classification of protein fold families. *Protein Eng.* **6**: 485–500.
- PARDEE, A. B., F. JACOB and J. MONOD, 1959 The genetic control and cytoplasmic expression of “inducibility” in the synthesis of  $\beta$ -galactosidase by *E. coli*. *J. Mol. Biol.* **1**: 165–178.
- PETRILLI, P., 1993 Classification of protein sequences by their dipeptide composition. *Comput. Appl. Biosci.* **9**: 205–209.
- PONTECORVO, G., 1950 New fields in the biochemical genetics of microorganisms. *Biochem. Soc. Symp.* **4**: 40–50.
- RAMBACH, A., 1990 New plate medium for facilitated differentiation of *Salmonella* spp. from *Proteus* spp. and other enteric bacteria. *Appl. Env. Microbiol.* **56**: 301–303.
- REEVES, P., 1993 Evolution of *Salmonella* O antigen variation by inter-specific gene transfer on a large scale. *Trends Genet.* **9**: 17–22.
- ROPER, J. A., 1950 Search for linkage between genes determining a vitamin requirement. *Nature* **166**: 956–957.
- ROSSMAN, M. G., D. MORAS and K. W. OLSEN, 1974 Chemical and biological evolution of a nucleotide-binding protein. *Nature* **250**: 194–199.
- ROTH, J. R., J. G. LAWRENCE, M. RUBENFIELD, S. KIEFFER-HIGGINS and G. M. CHURCH, 1993 Characterization of the cobalamin (vitamin B<sub>12</sub>) biosynthetic genes of *Salmonella typhimurium*. *J. Bacteriol.* **175**: 3303–3316.
- ROTH, J. R., N. BENSON, T. GALITSKI, K. HAACK, J. LAWRENCE *et al.*, 1996 Chromosomal rearrangements: formation and applications, pp. 2256–2276 in *Escherichia coli and Salmonella: Cellular and Molecular Biology*, Ed. 2, edited by F. C. NEIDHARDT, R. CURTISS III, J. L. INGRAHAM, E. C. C. LIN, K. BROOKS LOW *et al.* ASM Press, Washington, DC.
- SANDERSON, K. E., and M. DEMEREC, 1965 The linkage map of *Salmonella typhimurium*. *Genetics* **51**: 897–913.
- SANDERSON, K. E., A. HESSEL and K. E. RUDD, 1995 Genetic map of *Salmonella typhimurium*, Edition VIII. *Microbiol. Rev.* **59**: 241–303.
- SAXILD, H. H., and P. NYGAARD, 1988 Gene-enzyme relationships of the purine biosynthetic pathway in *Bacillus subtilis*. *Mol. Gen. Genet.* **211**: 160–167.
- SCHMIDT, J., M. BUBUNENKO and A. P. SUBRAMANIAN, 1993 A novel operon organization involving the genes for chorismate synthase (aromatic biosynthesis pathway) and ribosomal GTPase center proteins (L11, L1, L10, L12: *rplKAJL*) in cyanobacterium *Synechocystis* PCC 6803. *J. Biol. Chem.* **268**: 27447–27457.
- SELANDER, R. K., and B. R. LEVIN, 1980 Genetic diversity and structure in *Escherichia coli* populations. *Science* **210**: 545–547.
- SHANLEY, M. S., A. HARRISON, R. E. PARALES, G. KOWALCHUK, D. J. MITCHELL *et al.*, 1994 Unusual G+C content and codon usage in *catIJF*, a segment of the *ben-cat* supra-operonic cluster in the *Acinetobacter calcoaceticus* chromosome. *Gene* **138**: 59–65.
- SHARP, P. M., and W.-H. LI, 1987 The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**: 1281–1295.
- SLATIS, H. M., and D. A. WILLERMET, 1953 The miniature complex in *Drosophila melanogaster*. *Genetics* **39**: 45–58.
- SMITH, D. A., 1961 Some aspects of the genetics of methionineless mutants of *Salmonella typhimurium*. *J. Gen. Microbiol.* **24**: 335–353.
- SMITH, E. L., and E. MARGOLASH, 1964 Evolution of cytochrome *c*. *Fed. Proc.* **23**: 1243–1247.
- SPENSER, J. B., N. J. STOLOWICH, C. A. ROESSNER and A. I. SCOTT, 1993 The *Escherichia coli* *cysG* gene encodes the multifunctional protein, siroheme synthase. *FEBS Lett.* **335**: 57–60.
- STAHL, F. W., and N. E. MURRAY, 1966 The evolution of gene clusters and genetic circularity in microorganisms. *Genetics* **53**: 569–576.
- STEPHENS, S. G., 1951 Possible significance of duplications in evolution. *Adv. Genet.* **4**: 247–265.
- STEVENSON, G., B. NEAL, D. LIU, M. HOBBS, N. H. PACKER *et al.*, 1994 Structure of the O antigen of *Escherichia coli* K-12 and the sequence of its *rfb* cluster. *J. Bacteriol.* **176**: 4144–4156.
- STOKES, J. L., and H. G. BAYNE, 1958 Growth-factor-dependent strains of *Salmonellae*. *J. Bacteriol.* **76**: 417–421.
- SUBRAMANIAN, A. R., 1975 Copies of proteins L7 and L12 and heterogeneity of the large subunit of *Escherichia coli* ribosome. *J. Mol. Biol.* **95**: 1–8.
- SWANEN, M., 1994 Horizontal gene flow: evidence and possible consequences. *Annu. Rev. Genet.* **28**: 237–261.
- TITTENSOR, J., and J. R. ROTH, 1996 Unpublished results.
- VAN DE GUCHTE, M., J. KOK and G. VENEMA, 1991 Distance-dependent translational coupling and interference in *Lactococcus lactis*. *Mol. Gen. Genet.* **227**: 65–71.
- WHITFIELD, H. J., JR., D. L. GUTNICK, M. N. MARGOLIES, R. G. MARTIN, M. M. RECHLER *et al.*, 1970 Relative translation frequencies of the cistrons of the histidine operon. *J. Mol. Biol.* **49**: 245–249.
- WHITTAM, T. S., and S. AKE, 1992 Genetic polymorphisms and recombination in natural populations of *Escherichia coli*, pp. 223–246 in *Mechanisms of Molecular Evolution*, edited by N. TAKAHATA and A. G. CLARK. Japan Scientific Society Press, Tokyo.
- WHITTAM, T. S., H. OCHMAN and R. K. SELANDER, 1984 Geographical components of linkage disequilibrium in natural populations of *Escherichia coli*. *Mol. Biol. Evol.* **1**: 67–83.
- WITTMAN-LIEBOLD, B., A. K. E. KÖPKE, E. ARNDT, W. KRÖMER, T. HATAKEYAMA *et al.*, 1990 Sequence comparison and evolution of ribosomal proteins and their genes, pp. 598–616 in *The Ribosome*, edited by W. E. HILL, A. DAHLBERG, R. A. GARRETT, P. B. MOORE, D. SCHLESSINGER, *et al.* American Society for Microbiology, Washington, DC.
- WOEHLKE, G., K. WIFLING and P. DIMROTH, 1992 Sequence of the sodium ion pump oxaloacetate decarboxylase from *Salmonella typhimurium*. *J. Biol. Chem.* **267**: 22798–22803.
- XIANG, S. H., M. HOBBS and P. R. REEVES, 1994 Molecular analysis of the *rfb* gene cluster of a group D2 *Salmonella enterica* strain: evidence for its origin from an insertion sequence-mediated recombination event between group E and D1 strains. *J. Bacteriol.* **176**: 4357–4365.
- YANOFSKY, C., and E. S. LENNOX, 1959 Transduction and recombination study of linkage relationships among the genes controlling tryptophan synthesis in *Escherichia coli*. *Virology* **8**: 425–447.
- YURA, T., 1956 Evidence of nonidentical alleles in purine-requiring mutants of *Salmonella typhimurium*. *Carnegie Inst. Washington Publ.* **612**: 63–75.
- ZUCKERKANDL, E., and L. PAULING, 1962 Molecular disease, evolution, and genetic heterogeneity, pp. 189–225 in *Horizons in Biochemistry*, edited by M. KASHA and B. PULLMAN. Academic Press, New York.

Communicating editor: W.-H. Li